Identifying molecular markers for breeding a future oil crop, *Lepidium campestre*

Noomi Lodenius

Degree project • 60 credits • Second cycle
Swedish University of Agricultural Sciences, SLU
Department of Plant Breeding
Plant Biology for a sustainable production – Master Program –
Plant protection and Breeding for Mitigating Climate change
2022 - 2023

Identifying molecular markers for breeding a future oil crop, Lepidium campestre

Noomi Lodenius

Supervisor:	**Cecilia Hammenhag, Swedish University of Agricultural Sciences, SLU, Department of Plant Breeding**
Assistant supervisor:	Mulatu Geleta Dida, Swedish University of Agricultural Sciences, SLU, Department of Plant Breeding
Examiner:	Therése Bengtsson, Swedish University of Agricultural Sciences, SLU, Department of Plant Breeding

Credits:	60 credits
Level:	Second Cycle
Course title:	Independent project in Biology, A2E
Course code:	EX1001
Programme/education:	Plant Biology for a sustainable production – Master Program – Plant Protection and Breeding for Mitigating Climate Change
Course coordinating dept:	Department of Plant Breeding
Place of publication:	Epsilon, SLU
Year of publication:	2023
Cover picture:	(Private photos, Noomi Lodenius)
Copyright:	All images are used with permission from the copyright owner (Figure 2).
Keywords:	Glucosinolate content, interspecific hybridization, *Lepidium campestre*, marker discovery, oil content, oil quality

Swedish University of Agricultural Sciences
Faculty of Landscape Planning, Horticulture and Agricultural Sciences (LTV faculty)
Department of Plant Breeding

Abstract

The cold hardy novel oil- and cover crop *Lepidium campestre* is currently being domesticated as a potential oil crop for the Nordic region. To achieve this objective, multiple traits have previously been identified as desirable to improve, including seed oil content and composition, and glucosinolate (GL) content. This study aims to find improve future breeding efforts for these traits, by identifying favourable markers in *L. campestre*, as well as evaluate related species for future interspecific hybridization. In line with this, major genes involved in the biosynthetic and degradation pathways of seed oil and glucosinolate were evaluated in 40 accessions of *Lepidium* with the aim of finding a significant association with polymorphisms and total oil content, oleic acid (OA) and erucic acid (EA) contents as well as total GL content, Sinalbin (Sb) and glucoallysin (Gl) contents. In total, 113 significantly associated markers were identified. Among these markers, 27 were identified as especially interesting, 13 markers with oil content, eight with oil composition (OA and EA), and six with GL content. In addition, relatives of *L. campestre* were evaluated for morphology, seed oil content, oil composition and glucosinolate content. Their phylogenetic relationship with *L. campestre* was also evaluated for use as potential candidates for interspecific hybridization. In this analysis two species with beneficial OA content were suggested for future crosses, *L. hirtum subsp. calycotrichum* and *L. heterohyllum*. An additional four species with beneficial oil-, OA-, EA- and GL-content were identified as interesting targets for future embryo rescue protocol adaptation, necessary to overcome breeding barriers. These include *L. graminifolium, L. sativum, L. virginicum subsp. Menziesii,* and *L. perfoliatum.*

Keywords: Glucosinolate content, interspecific hybridization, *Lepidium campestre*, marker discovery, oil content, oil quality

Contents

List of tables **7**

List of figures **10**

Abbreviations **13**

1. Introduction **15**

1.1 Background and Aim 15

1.2 A brief history and current state of oil crop domestication 16

1.3 Lepidium campestre 17

1.3.1 *L. campestre* as a potential oil crop 17

1.3.2 Studied oil content, composition and glucosinolate-related genes in *L. campestre* 18

1.4 Oil content and fatty acid composition of *L. campestre* 19

1.5 Glucosinolates in *L. campestre* 21

1.6 DNA sequencing methods 24

1.7 Genetic polymorphisms 25

1.8 Promoting *L. campestre* domestication through interspecific hybridization 27

2. Material & Method **29**

2.1 Plant material 29

2.2 Bioinformatic analysis 29

2.2.1 Gene structure predictions in *L. campestre* 29

2.2.2 Bioinformatic study of Brassicaceae species 30

2.3 Polymorphism discovery in *L. campestre* genes 30

2.3.1 DNA Extraction 31

2.3.2 PCR and Primer design 31

2.3.3 Sequencing 32

2.3.4 Statistical analysis on genetic variation 33

2.4 RNA analysis of *PEN2* 33

2.4.1 RNA Extraction and cDNA synthesis 33

2.4.2 PCR and Primer design for amplifying RNA 34

2.5 Characterization of *Lepidium* species 34

2.5.1 Analysis of agronomically important traits 34

2.5.2 *Lepidium* seed oil extraction 35

2.5.3 Evaluating the effect of selected polymorphisms in *FAD2*, *LEC2*, *WRI1* and *TAG1* in *Lepidium* species 36
2.6 Sequencing method analysis 36

3. Results 38
3.1 Bioinformatic analysis of *FAD2*, *FAE1*, *PEN2* & *SOT16* in selected Brassicaceae species 38
3.1.1 *FAD2* 38
3.1.2 *FAE1* 40
3.1.3 *PEN2* 42
3.1.4 *SOT16* 45
3.2 Evaluating polymorphisms found in *L. campestre* 47
3.2.1 Polymorphisms identified through Sanger sequencing 47
3.2.2 Polymorphisms identified using targeted NGS 48
3.2.3 Significantly associated polymorphisms to target traits found using targeted NGS sequencing 51
3.2.4 Principle Component Analysis (PCA) 54
3.3 Method comparison – Targeted NGS and Sanger sequencing 56
3.4 Evaluation of traits in *Lepidium* species for future interspecific breeding with *Lepidium campestre* 58
3.4.1 Evaluation of selected polymorphisms in oil content related genes in *Lepidium* species 58
3.4.2 OiI content and oil composition analysis in *Lepidium* species 68
3.4.3 Lepidium species for future interspecific breeding with *L. campestre* 69
3.4.4 Glucosinolate content and composition in Lepidium species 71
3.4.5 Germination test in *Lepidium* species 72
The development, morphology, flower induction and seed-setting in *Lepidium* species 74
Summary of evaluation of *Lepidium* species for future interspecific hybridization . 78

4. Discusion 79
4.1 SNP discovery in *L. campestre* 79
4.2 Future approaches to improve key traits in *L. campestre* 81
4.3 Bioinformatics 82
4.3.1 In Brassicaceae 82
4.3.2 Sequence analysis of targeted polymorphisms in *FAD2*, *WRI1*, *TAG1* and *LEC2* in *Lepidium* 84
4.4 Evaluation for future interspecific hybridization 85
4.5 Conclusion 87

References 89

Popular science summary 104

Acknowledgements 106

Appendix 1 ..... **107**

Appendix 2 ..... **108**

Appendix 3 ..... **109**

Appendix 4 ..... **111**

Appendix 5 ..... **113**

Appendix 6 ..... **114**

List of tables

Table 1. Seed oil content and major fatty acids present ≥ 4 % of total oil content in *L. campestre*, compared to *B. napus*. 19

Table 2. Morphology, known traits, and general habitat for *Lepidium* species evaluated in this study for future interspecific hybridization with *L. campestre*. Species with multiple reported chromosome numbers are described as 2n = X, Y. 28

Table 3. List of genes used for genetic variation analysis in this study. Coding sequence predicted from *L. campestre* genomic sequence after alignment with *A. thaliana* CDS. Full name and function based on data gathered from The Arabidopsis Information Resource (TAIR, 2023) based on functional studies in *A. thaliana.* CDS = Coding sequence, VLCFAs = Very long chain fatty acids. 30

Table 4. The six unique polymorphisms found across the four studied genes using Sanger sequencing. POS refers to position in its linkage group, and POS CDS to its bp position counted from ATG. A '+' refers to a polymorphism found downstream of CDS. A '*' in alternative allele refers to an ambiguous base, Y = C/T, M = A/C. Abbreviations: Alt. – Alternative, OA - Oleic acid, GLs - Glucosinolates, sb - sinalbin. Ref. – reference. 47

Table 5. The polymorphisms found across 15 studied genes (excluding accession LEP2020-313 and the polymorphism only recorded for this accession), shown as the number of polymorphisms in each gene as total, up-and downstream of the gene, introns and exons. 48

Table 6. List of missense mutations and codon deletions found in oil- content and quality related genes of *L. campestre*. POS LG refers to its bp position in the linkage group, POS CDS to its bp position from ATG in the CDS. Codon refers to which codon is then affected counted from ATG as codon 1. Exon 7 * (TAG1) indicated that the exon is predicted with low support. Polyphen scores evaluate the probability of a substitution being damaging for protein. Values between 0-0.15 are predicted to be benign, 0.15-0.85 possibly damaging and 0.85-1 probably damaging. 50

Table 7. A list of all 113 polymorphisms observed in at least two accessions, with significant associations to traits of interest (OC, OA, EA, GL, Sb and Gla) in 39 studied accessions..... 53

Table 8. Distribution of marker groups with significant correlation to oil (A) or glucosinolate content (B), for each accession. No. In Group indicate the number of polymorphic loci in a group (colour coded as in Table 7, column "name of polymorphism") which are shared by the same accessions. Trait association indicate which trait the marker is significant for, while R indicates that reference allele is preferred and A that alternative allele is preferred. 54

Table 9. Cost (in SEK) for sequencing and identification of polymorphisms in the 15 genes included in this study by the targeted Next-Generation Sequencing (NGS) (CD-genomics), and the theoretical cost of using Sanger sequencing (at Eurofins) for purified Polymer Chain Reaction (PCR) products in tubes, plate or unpurified products in plate format. Costs for Sanger sequencing were calculated by estimates based on data from 15 genes and 40 accessions in this study with average gene length of 2080 bp +-700 bp up and downstream and five primer pairs per gene, including cost of PCR and sequencing-related products and services. Polymorphism refers to a single locus with a polymorphism in at least one accession (Table 7). Exchange rate used was 1 USD = 10.40 SEK. 57

Table 10. Polymorphisms discovered in L. campestre (Table 7) chosen for further evaluation in Lepidium accessions. Phenotypes marked in red indicates negative correlation, while green indicates positive. POS indicates the position of locus in relationship to the coding sequence, with – being upstream of ATG and + being downstream of stop-codon. 58

Table 11. Table of Lepidium species identified with either chromosome number 2n =16 and/or close phylogenetic relationship. Chromosome number & country of origin & alternative name from [a]Warwick & Al-Shehbaz (2006) if not otherwise stated. Relatedness was determined from phylogenetic trees of [b]Mummenhoff *et al.* (2009), [c]Mummenhoff, Brüggemann & Bowman (2001) and [d]Ma *et al.* (2020). Current gene bank availability for distribution is a summary of [e]Genesys (2023) and [f]Kew (2023a) and verified from distributing gene bank webpages. Chromosome number (A) or (B) corresponds to findings in country (A) or (B) in country of origin where chromosome number may be tied to location. Relatedness: Green = Very close within 1-2 nodes (1), Yellow = close, within 3 nodes (2). Brown = Distant, more than 4 nodes (3). Grey = Unknown (4). 69

Table 12. Literature study of Lepidum accessions regarding glucosinolate content and composition. The very low GLs content reported in *L. graminifolium* (*) refers to a *L. iberis* accession. Reported values have been converted to µmol/g dry seed weight. Subspecies for *L. hirtum* are not included since no study has evaluated their GL content. 72

Table 13. Morphology types (see Figure 20), plant height, phenology, and seed- weight and size for the 16 studied Lepidium accessions. Morphology types notes as

"N/A" for species where the trait could not be examined due to time constraint. TSW and seed size n = 48 – 1642, plant height at flower induction n = 2-5. Abbreviations: TSW – thousand seed weight. 74

Table 14. Simplified summary of figure 18-19 and Table 11-13, to showcase potential of the 15 Lepidium accessions evaluated in this study. Traits designated as medium-level are comparable to those of *L. campestre* while good/very good is better and poor/very poor/abysmal is worse. Medium/poor phylogenetic distance refers to crossing barrier. Column 2n = 16 "yes" refers to the accession having only the ideal chromosome number of 2n = 16, and yes/no that accession are reported to have both 2n =16 and additionally reported chromosome numbers of 2n ≠ 16 which may act as a crossing barrier. 78

List of figures

Figure 1. Flowchart showcasing study aim, process, and methods of the two projects contained within this study. The first project evaluating polymorphisms to identify molecular markers for future breeding efforts is shown in blue/purple. The second project evaluating relatives of *L. campestre* to identify targets for future interspecific hybridization is shown in orange/pink/brown. Bioinformatic-related work is shown in green. 15

Figure 2. Young and leafy *L. campestre* plant (left) (photo Noomi Lodenius) and matured flowering plants with pods in a field trial in Alnarp (right) (Gustafsson, 2018).. 17

Figure 3. Simplified FA-synthesis, elongation and modification scheme featuring genes of interest for this study in **bold**. Fatty acids are synthesized in the plastids and transported to the ER for TAG-assembly. Master regulators of oil biosynthesis are located in the nucleus. 20

Figure 4. Simplified GL-synthesis schematic representation, featuring genes of interest in this study in **bold**. The figure is based on the work of Harun *et al.* (2020), Reintanz *et al.* (2001) and Gigolashvili *et al.* (2007). GL – Glucosinolates, TF – Transcription factor. 22

Figure 5. Schematic illustration of SNP classification and predicted impact on protein structure and function. 26

Figure 6. Phusion polymerase PCR program. 32

Figure 7. Visualization of *FAD2* gene conservation and sequence similarity across selected Brassicaceae species compared to *L. campestre.* Orange indicates a polymorphism, blue indicate an insertion. Regions of conserved histidine cluster motifs are marked with yellow triangles. ps – pseudogene, chr. – Chromosome 39

Figure 8. Visualization of *FAE1* gene conservation and sequence similarity across selected Brassicaceae species compared to *L. campestre*. Orange indicates a polymorphism, blue indicate an insertion. 41

Figure 9. PCR-primer products for mRNA-characterization (B) aligned to the predicted mRNA product of L. campestre *PEN2* (A). Prediction based on *A. thaliana PEN2* structure. A PCR-product was obtained and verified on gel for the region

spanning exon four to five, but not sequenced due to package handling issues, and hence not shown. Abbreviations: Ex = exon, F = forward primer, R = reverse primer. 42

Figure 10. Visualization of *PEN2*-conservation across selected Brassicaceae species compared to *L. campestre*. *L. campestre PEN2*-Exon (E)/ Intron (I) gene structure based on CDS-sequences from *A. thaliana (*TAIR, 2023) supported by experimental data (Figure 9). Orange indicates a polymorphism, and blue indicates an insertion. 44

Figure 11. Visualization of *SOT16* gene conservation and sequence similarity across selected Brassicaceae species compared to *L. campestre.* Orange indicates a polymorphism, and blue indicates an insertion. 46

Figure 12. Principle component analysis (PCA) of polymorphisms identified in *L. campestre*. 55

Figure 13. Visualization of *FAD2* gene alignment for 12 *Lepidium* species in positions 28-519 bp. A band in the *L. campestre* sequence indicates where an interesting polymorphism has been found in *L. campestre*, an orange band in the *Lepidium* accessions indicates substitutions compared to *L. campestre*. Region around the target polymorphism (*FAD2*_16852827-C+2/-) shown in detail, in blue. A yellow arrow indicates where a conserved motif is located. 59

Figure 14. (A) - Visualization of the predicted *WRI1* structure based on Arabidopsis TAIR data, with the three sequenced amplicons shown in green. Gene alignment in the studied regions between -325-625 bp (Amplicon 1) (B), 903-1826 bp (Amplicon 2) (C) and 3555-+213 bp (Amplicon 3) (D), position relative to CDS from ATG. A band in the *L. campestre* sequence indicates an interesting polymorphism, an orange band in the *Lepidium* accessions indicates substitutions compared to *L. campestre*, and a blue band an insertion with an arrow for longer insertions. A yellow arrow indicates where a conserved domain is located. Sequence similarity is not included when amplicons are <50% coverage to reference. Regions around the target polymorphisms (*WRI1*_16471364-C/C+1,C+2, *WRI1*_16472916-G/G+1, *WRI1*_16473045-C/T, *WRI1*_16473352-C/C+7 and *WRI1*_16475686-T/C) shown in detail, in blue. 62

Figure 15. (A) - Visualization of the predicted *TAG1* structure based on Arabidopsis TAIR data, with the two sequenced amplicons shown in green. Gene alignment in the studied regions between 578-961 bp (Amplicon 1) (B) and 2996-+545 bp (Amplicon 2) (C), position relative to CDS from ATG. A band in *L. campestre* indicates an interesting SNP, an orange band in the *Lepidium* accessions indicates substitutions compared to *L. campestre*, a blue band an insertion with an arrow for longer insertions Sequence similarity is not included when

amplicons are <50% coverage to reference. Regions around the target polymorphisms (*TAG1*_602057-A/G, *TAG1*_604746-C/T and *TAG1*_604920-T/G) shown in detail, in blue. 64

Figure 16. Visualization of the *LEC2* gene alignment in the studied regions between 2388-3365 bp, position relative to CDS from ATG. A band in *L. campestre* indicates where an interesting polymorphism has been found, an orange band in the Lepidium accessions indicates substitutions compared to *L. campestre,* a blue band an insertion with an arrow for longer insertions. Sequence similarity is not included for amplicons with <30% coverage of query sequence. Regions around the target polymorphisms (*LEC2*_3905801-T+2/-, *LEC2*_3905804-C/T, *LEC2*_3905844-A+2/-, *LEC2*_3905849-C/T and *LEC2*_3906180-A/G) shown in detail, in blue. 66

Figure 17. Phylogenetic tree of 15 *Lepidium* accessions based on *FAD2* amplicon (Figure 13) sequences generated with neighbour joining with *B. napus* as an outgroup. Each branch length distance is displayed, with higher numbers indicating greater phylogenetic distance. 67

Figure 18. Oil composition in 16 *Lepidium* accessions. n = 10x3 (seeds), except *L. ruderale* and *L. hirtum subsp. atlanticum* with n = 10x2. Sources for previously reported oil content: [a]Nilsson, Johansson & Merker (1998), [b]Kjaer *et al.* (1954) 69

Figure 19. Germination rate (%) of Lepidium species at day 4, 8 and 12 after plating seeds (n=50) in a petri dish with sterile water and filter paper and survival rate (% of seeds germinated and survived), of seedlings (n=10) after being placed in soil, and carefully watered three times a week for 38 days. 73

Figure 20. Observed and sketched morphological variation among *Lepidium* species in this study. 77

Abbreviations

Acc.	Accession
AGG	Australian Grains Genebank
BLAST	Basic Local Alignment Tool
Bp	Base pair
BUS1	*BUSHY 1*
CDS	Coding sequence
cDNA	Complementary DNA
CIP	International Potato Center
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
CTAB	Hexadecyltrimethylammonium Bromide
DAG	Diacylglycerol
DNA	Deoxyribonucleic Acid
EA	Erucic Acid
EMS	Ethyl Methanesulfonate
ER	Endoplasmic Reticulum
FAD2	*FATTY ACID DESATURASE 2*
FAE1	*FATTY ACID ELONGASE 1*
Gla	Glucosallyssin
GL	Glucosinolates
GM	Genetic Modification
GTR1, 2	*GLUCOSINOLATE TRANSPORTER 1, 2*
INDELS	Insertions/Deletions
IPK	Genebank of Leibniz-Institute of Plant Genetics and Crop Plant Research Gatersleben
KCS8	*3-KETOACYL-COA-SYNTHASE 8*
LA	Linolenic Acid
LARI	Lebanese Agricultural Research Institute
LEC1, 2	*LEAFY COTYLODON 1, 2*
MAMP	Microbe-Associated Molecular Pattern
MAS	Marker-assisted Selection

mRNA	Messenger RNA
MSB	Millennium Seed Bank Kew Gardens
MYB28	*MYB DOMAIN PROTEIN 28*
NCBI	National Center for Biotechnology Information
NGS	Next Generation Sequencing
NPGS	USDA National Plant Germplasm System
Nt	Nucleotide
OA	Oleic Acid
OC	Oil Content
ORF	Open Reading Frame
PCR	Polymerase Chain Reaction
PEN2	*PENETRATION 2*
RNA	Ribonucleic Acid
ROD1, 2	*REDUCED OLEATE DESATURATION 1, 2*
SLU	Swedish University of Agricultural Sciences
Sb	Sinalbin
SOT16	*SULFOTRANSFERASE 16*
TAG	Triacylglycerol
TAG1	*TRIACYLGLYCEROL 1*
TAIR	The *Arabidopsis* Information Resource
TF	Transcription Factor
UPM-BGV	The Plant Germplasm Bank of the Polytechnic University of Madrid
VIR	Nikolaj Ivanovitj Vavilov All-Russian Institute of Plant Genetic Resources
WRI1	*WRINKLED 1*

1. Introduction

1.1 Background and Aim

One of the major challenges we face in moving towards a sustainable future is the demand for oil. Oil is most commonly derived from non-renewable sources, affecting the climate negatively by being a major factor in carbon emissions. Furthermore, the demand for edible oil has tripled in the last two decades (Jindasa *et al.* 2022). Vegetable oils share high structural similarity with fossil fuels and provide a much more sustainable alternative (Carlsson *et al.* 2011). Rapeseed (*Brassica napus ssp. napus*) is the major oil crop grown in Sweden. It is, however, not very productive in the Nordic climate, due to its poor winter hardiness. For this purpose, domestication efforts of a cold hardy field cress (*Lepidium campestre*), a wild member of the Brassicaceae family, have been put forward in the last decades (Nilsson, Johansson & Merker, 1998, Gustafsson *et al.* 2018, Ortiz *et al.* 2020).

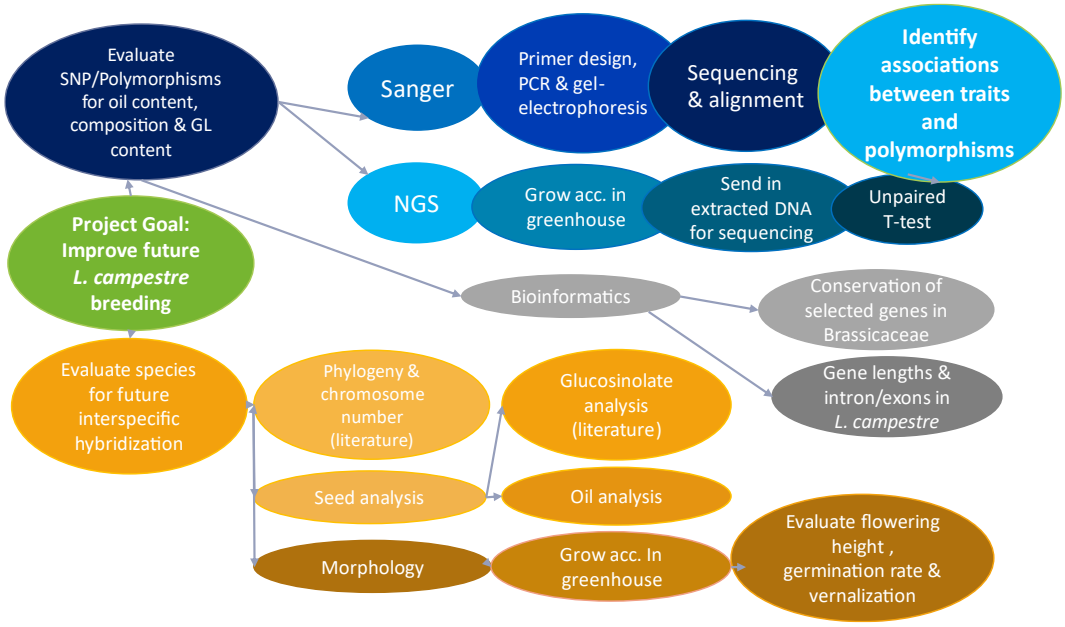


Figure 1. Flowchart showcasing study aim, process, and methods of the two projects contained within this study. The first project evaluating polymorphisms to identify molecular markers within *L. campestre* for future breeding efforts is shown in blue. The second project evaluating relatives of *L. campestre* to identify targets for future interspecific hybridization is shown in orange. Bioinformatic-related work is shown in grey.

Acc = Accession, NGS = Next generation sequencing, PCR = Polymer Chain Reaction.

The primary objective of this study is to identify molecular markers and to evaluate and identify potential new *Lepidium* species that could be used for interspecific hybridization with *L. campestre*, thereby introgressing desirable genes to the latter to speed up its domestication and breeding (Figure 1). In order to identify markers, genetic variation analysis was performed on *L. campestre* genes related to oil content, oil composition, glucosinolate biosynthesis and degradation in order to identify polymorphism that can serve as future molecular markers for the improvement of *L. campestre.* Relatives in the *Lepidium* genus were then evaluated for beneficial traits, such as oil content, oil composition, seed weight and seed germination rate, for future interspecific hybridization with *L. campestre.*

1.2 A brief history and current state of oil crop domestication

Humans have cultivated crops for the past 13 000 years, with a wide establishment around 5000-7000 BCE (Tauger, 2010). Many oil crops that are culturally important today have been cultivated since ancient times. Among the early oil crops are olives, sesame, and safflower, with cultivation dating back to 2000-2500 BCE (Lanza, 2011, Dorian, 2003, Smith, 1996). The most traditional forms of crop improvement include selecting plants based on visual observable phenotypic traits of interest, which is still used today (Meyer et al, 2012). Over time, plant breeding methods have grown more sophisticated. Mendel's discovery of the laws of inheritance in the 1800's has provided essential tools for understanding inheritance of traits. Discovery and understanding of the genetic code further allowed for implementation of molecular markers and genetic engineering in the 1980-1990's (Schlötterer, 2004, Prado *et al.* 2014). These methods as well as marker-assisted selection (MAS) and genomic selection used extensively today, have greatly improved the speed of trait selection and enhancement. In addition, the current decade has seen the rise of more precise genetic engineering methods, such as Clustered Regularly Interspaced Short Palindromic Repeats-CRISPR-associated protein 9 (CRISPR-Cas 9) (Ma, Zhang & Huang, 2014).

Although oil crop domestication has a longstanding history, only a few potential oil crops are currently under domestication due to the long time and high costs associated with domestication (Ortiz *et al.* 2020). Emerging future oil crops include *Brassica carinata, Eruca sativa, Camelina sativa* and *Crambe spp.*, the last two suitable for northern climate however with domestication efforts focused on breeding for industrial purposes (Farooq *et al.* 2015). Due to emerging needs to increase plant oil production in both industry and edible oils, *L. campestre* was chosen for domestication (Andersson *et al.* 1999, Ortiz *et al.* 2020).

1.3 Lepidium campestre

1.3.1 *L. campestre* as a potential oil crop

Field cress (*L. campestre*) (Figure 2) has multiple desirable characteristics contributing to its potential as an oil crop. It is a diploid biennial self-fertilized plant from the Brassicaceae family with a relatively small genome and a chromosome number of 2n =16 (Geleta *et al.* 2020). It is naturally high yielding, winter hardy with an upright growth habit (Andersson *et al.* 1999, Nilsson, Johansson & Merker, 1998, Ivarsson *et al.* 2016). Due to its biennial nature, it is suitable for growing alongside other crops in the field as a catch or cover-crop, enabling higher efficiency of land use, soil retention and reducing NO_3 leaching into the soil (Ivarsson *et al.* 2016). It is native to most of Europe with a range from Turkey in the south to Sweden in the north, and it has been introduced to North America, South Africa, Japan, China, and Australia (Kew, 2023B). Efforts are currently underway to advance the domestication of field cress and to breed for higher oil content, more desirable oil composition and perenniality (Gustafsson *et al.* 2018, Ortiz *et al.* 2020, Ivarsson *et al.* 2016). The relatively close phylogenetic relationship between *L. campestre* and the model species *Arabidopsis thaliana* as well as the widely cultivated oil crop *B. napus* further facilitates the research and understanding of its genomic, transcriptomic, and proteomic makeup. Whole genome sequencing of field cress (NCBI, 2023) and additional targeted sequencing efforts have enabled the identification of molecular markers for beneficial traits. Additional genomic tools and resources have been developed, such as the identification of Quantitative trait loci (QTL) for multiple desirable traits, including plant height, seed yield and pod shattering resistance (Hammenhag *et al.* 2020, Geleta *et al.* 2020).

Figure 2. Young and leafy *L. campestre* plant (left) (photo Noomi Lodenius) and matured flowering plants with pods in a field trial in Alnarp (right) (Gustafsson, 2018).

L. campestre seed oil composition has been found to be suitable for certain industrial purposes (Gustafsson *et al.* 2020). However, its high levels of erucic acid (EA) hinder the use of *L. campestre* oil in the food industry, as EA is not easily digestible by humans and causes toxicity (Vles, Bijster & Timmer, 1978). After the oil has been pressed or extracted, the remaining seed cake can be used as animal feed. The high glucosinolate content in *L. campestre* seeds (Arefaine *et al.* 2019), however, is an issue for this application, since its breakdown products are toxic to livestock, especially for non-ruminant animals such as pigs (Tripathi & Mishra 2007, Burel *et al.* 2000).

1.3.2 Studied oil content, composition and glucosinolate-related genes in *L. campestre*

Major genes known to affect oil content, oil composition and glucosinolate levels have been identified in *L. campestre* through comparative genomics (Gustafsson 2018). Some of these genes have been targeted for genetic modification, validating their function, such as *FATTY ACID DESATURASE-2* (*FAD2*), *FATTY ACID ELONGASE-1* (*FAE1*), *REDUCED OLEATE DESATURATION 1* (*ROD1*) and *WRINKLED-1* (*WRI1*) (Ivarsson *et al.* 2016, Ivarsson *et al.* 2017, Sandgrind *et al.* 2023). Genes known to regulate glucosinolate levels in seeds, including *GLUCOSINOLATE TRANSPORTER-1* and *2* (*GTR1* and *GTR2*) have also been targeted for genetic engineering (Sandgrind *et al.* 2022) with a near complete abolishment of GL content in seeds as a result.

In this study, the *L. campestre* genes related to oil composition, *FAD2* and *FAE1*, and glucosinolate biosynthesis and degradation-related genes, *PENETRATION-2* (*PEN2*) and *SULOTRANSFERASE-16* (*SOT16*), were evaluated for polymorphism. In addition, a genetic variation analysis was also performed in order to identify potential polymorphisms in *LEAFY COTYLODON-1* and 2, *LEC1*, *LEC2*, *TRIAGLYCEROL-1* (*TAG1*) and *WRI1* known to regulate oil content, *FAE1_1*, *FAE1_2*, *FAE1_3*, *FAD2* and *3-KETOACYL-COA-SYNTHASE-8* (*KCS8*) known to affect the fatty acid composition, and *BUSHY-1* (*BUS1*), *SOT16*, *PEN2*, *GTR1*, *GTR2* and *MYB DOMAIN PROTEIN-28* (*MYB28*) known to regulate glucosinolate content in seeds.

1.4 Oil content and fatty acid composition of *L. campestre*

Previous research has shown that *L. campestre* seeds contain 12-20% oil (Nilsson, Johansson & Merker, 1998, Geleta *et al.* 2020) and the seed oil is composed of 34-35% linolenic acid (LA; 18:3), 22-34% erucic acid (EA; 22:1), 15-16% oleic acid (OA; 18:1), as well as 9-11% Linoleic (18:2), 5-6% Eicosenoic (20:1) and 4-5% Palmitic acid (16:0) (Table 1).

Table 1. Seed oil content and major fatty acids present ≥ 4 % of total oil content in *L. campestre*, compared to *B. napus.*

Species	Seed oil content	Linolenic acid % (18:3)	Erucic acid % (22:1)	Oleic acid % (18:1)	Linoleic acid % (18:2)	Eicosenoic acid % (20:1)	Palmitic acid % (16:0)
L. campestre (field cress)	12-20[b, c]	34–35[a]	22-34 [a]	15-16[a]	9-11[a]	5-6[a]	4–5[a]
B. napus (rapeseed, low GL cultivar)	44-50[d]	10[e]	<0.5[e]	62[e]	22[e]	-	4[e]

[a]Andersson *et al.* (1999), [b]Nilsson, Johansson & Merker (1998), [c]Geleta *et al.* (2020), [d]Barthet & Daun (2011), [e]Kerr & Dunford (2018).

The oil composition of a crop can be altered through plant breeding. To decrease the toxicity of *B. napus* seed oil, low erucic acid cultivars have been developed (Wang *et al.* 2022, Stefansson & Hougen, 1964). However multiple projects have also been successful in increasing erucic acid content for the oleochemical industry, including ultra-high erucic acid oil *Crambe abyssinica* (Li *et al.* 2012) and high erucic acid rapeseed oil (Nath, 2009).

Although nutritionally comparable, LA is 10-40 times more susceptible to oxidation than OA, posing a problem for food oil shelf life and biofuel production (Cao *et al.* 2021, Dar *et al.* 2017). Therefore, decreasing linolenic and erucic acid in favor of their precursor oleic acid is a target in *L. campestre* breeding.

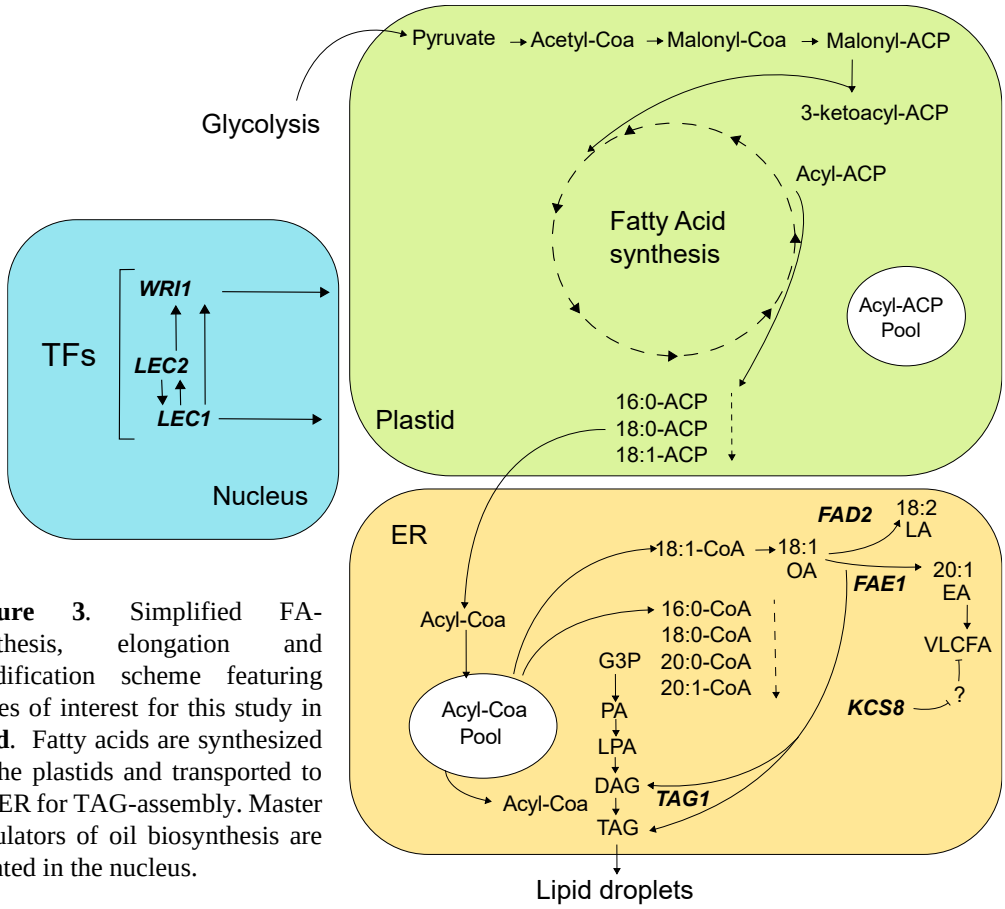


Figure 3. Simplified FA-synthesis, elongation and modification scheme featuring genes of interest for this study in **bold**. Fatty acids are synthesized in the plastids and transported to the ER for TAG-assembly. Master regulators of oil biosynthesis are located in the nucleus.

ACP = Acyl carrier protein, CoA = Coenzyme A, DAG = diacylglycerol, EA = Erucic acid, ER = Endoplasmic reticulum, FA = Fatty acid, G3P = Glyceraldehyde 3-phosphate, LA = Linoleic acid, LPA = Lysophosphatidic acid, OA = oleic acid, PA = Phosphatidic acid, TAG = Triacylglycerol, TF = Transcription factor, VLCFA = Very long chain fatty acid. The figure is based on the work of Bowsher, Steer & Tobin (2008), Kumar (2020) (TFs), Li-Beisson *et al.* (2013) (*FAD2*), Snell (2019) (general pathway), Park *et al.* (2022) (*FAD2*, FAE), Srinivas *et al.* (2012) (*FAE1*), Svatoš et al (2020) (TFs), Zhukov & Popov (2022) (*KCS8*).

Oleic acid is synthesized in plastids (Somerville & Browse, 1991). Linolenic acid biosynthesis can then occur after transportation to the ER when monounsaturated oleic acid (OA; $18:1^{\Delta 9}$) becomes desaturated by *FAD2*, allowing the formation of a second double bond at Δ12 (Miquel & Browse, 1992) (Figure 3). The *FAD2* gene contains multiple conserved histidine-rich iron-binding motifs essential for reduction and subsequent gene function (Cao *et al.* 2021). Transgenic low erucic/high oleic acid genotypes have been achieved by targeting *FAD2* in *L. campestre* with CRISPR-Cas9 (Sandgrind *et al.* 2023), and RNA-interference (RNAi) silencing (Ivarsson *et al.* 2016). *FAD2* has therefore been chosen for its potential in reducing linoleic acid in favour of its more beneficial precursor oleic acid.

TAG1 (also known as *ACYL-COA DIACYLGLYCEROL ACYLTRANSFERASE 1* (*DGAT*)) is a key enzyme for the triacylglycerol (TAG) accumulation in plant tissue (Colette *et al.* 2001), mediating the last step between diacylglycerol (DAG) and TAG. Upregulation of *TAG1* has been shown to greatly enhance seed oil content (Colette *et al.* 2001), while downregulation only somewhat decreases it (Kun *et al.* 2017). It may therefore be useful for enhancing oil seed content.

The condensing enzyme *FAE1* is of great interest for *L. campestre* erucic acid reduction, as it since long has been the target in breeding for obtaining low erucic acid lines and cultivars in *B. napus* (James *et al.* 1995, Wang *et al.* 2008). The enzyme is responsible for fatty acid (FA) chain elongation from oleic acid ($18:1^{\Delta 9}$) to erucic acid ($20:1^{\Delta 9}$), enabling one of the pathways for the synthesis of very long chain fatty acids (VLCFAs) (Wang *et al.* 2022, Zhukov & Popov, 2022) (Figure 3). *FAE1* belongs to one of the four VLCFA elongase complex enzyme groups, the rate-limiting KCS (Zhukov & Popov, 2022). The KCS-enzymes perform an initial condensation step, followed by a reduction by the Ketoacyl-Coenzyme A Reductase (KCR) group, a dehydration by the 3-hydroxy acyl-CoA dehydratase (HCD) group and a final reduction by the enoyl-CoA reductase (ECR) group, resulting in a FA that is two carbons longer (Zhukov & Popov, 2022. *KETOACYL-COA SYNTHASE 8* (*KCS8*) also belongs to the KCR group and studies indicate that this enzyme enhances the accumulation of VLCFAs through suppression of negative gene regulators (Zhukov & Popov, 2022).

The genes *WRI1*, *LEC1* and *LEC2* are considered master regulators of fatty acid biosynthesis with multiple targets within the pathway (Bowsher, Steer & Tobin 2008, Kumar 2020) (Figure 3). Loss of function mutations in *WRI1* have been demonstrated to cause up to 80% decrease in TAG accumulation in *A. thaliana* (Focks & Benning 1998). Work in *L. campestre* where *WRI1* was overexpressed through introduction of a new gene copy from *A. thaliana*, was shown to increase seed TAG by 29% (Ivarsson *et al.* 2017). Transcriptional level analysis of *WRI1* expression suggests that *LEC1* and *LEC2* act upstream of *WRI1*, influencing its expression (Kong, Yuan & Ma, 2019, Pelletier *et al.* 2017). Due to their key roles in oil biosynthesis, *WRI1*, *LEC1* and *LEC2* are all interesting targets in *L. campestre* breeding for increased seed oil content.

1.5 Glucosinolates in *L. campestre*

Decreased glucosinolate content is a major breeding target for *L. campestre.* Like other members of the Brassicaceae family, *L. campestre* has glucosinolates (GL) in different plant parts, such as leaves and seeds. This group of compounds is responsible for the characteristic sharp flavor of the Brassicaceae family. The glucosinolates are biologically inactive. However, disruption of cells through

wounding or herbivory causes myrosinase enzymes to release and break the glucosinolates down into a range of toxic breakdown products, thereby protecting the plants from herbivores (Rask *et al.* 2000). The insect herbivory-defense is enhanced greatly in the breakdown-products compared to their precursors, but so is their toxicity (Hopkins, Van Dam & Van Loon, 2009, Wittock & Burow, 2010). High glucosinolate seedcake used as feed can be detrimental to livestock health, including lowered production, and damage to the kidney(s) and liver (Tripathi & Mishra 2007, Burel *et al.* 2000). High glucosinolate feed, most notably from the Brassicaceae, therefore, requires expensive pre-treatment with high temperatures or microbial fermentation before use (Tripathi & Mishra 2007). Consequently, high glucosinolate levels in oil crops restrict the sustainability and profits of using seed oil cakes for animal feed.

However, as indicated above, glucosinolates do also have beneficial properties with regard to plant protection, including use for biofumigation, reducing soilborne pests and weeds (Gimsing & Kirkegaard, 2009) as well as resistance against insect herbivory (Hopkins, Van Dam & Van Loon, 2009). Recent research has also raised attention to their cancer chemoprotective properties (Martinez-Ballesta & Carvajal 2015).

Glucosinolates are secondary plant metabolites that fall into three groups based on what their amino acid side chains are synthesized from: aliphatic (mainly methionine), aromatic (phenylalanine, tyrosine) and indolic (tryptophan) (Figure 4). Glucosinolate biosynthesis can, in short, be described in three steps – i) side chain elongation, ii) core structure synthesis and iii) side chain modification (Harun *et al.* 2020). Total glucosinolate content varies drastically with age (Fahey, Zalcmann & Talalay 2001) and between plant tissues (Anderson *et al.* 2013).

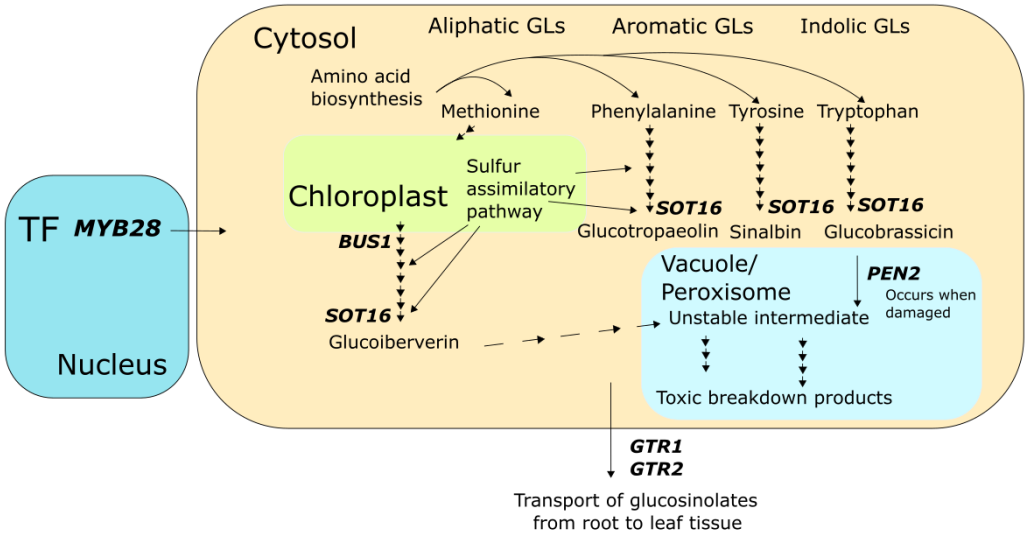


Figure 4. Simplified GL-synthesis schematic representation, featuring genes of interest in this study in **bold**. The figure is based on the work of Harun *et al.* (2020), Reintanz *et al.* (2001) and Gigolashvili *et al.* (2007). GL – Glucosinolates, TF – Transcription factor.

The most abundant glucosinolate in field cress is 4-hydroxybenzylglucosinolate (sinalbin) and the overall glucosinolate content in the seed ranges from typically 125-138 µmol/g (Arefaine *et al.* 2019) up to 600 µmol/g (Isoz, 2018). For reference, European Union glucosinolate limit in animal feed is set at 30 µmol/g (Europeiska kommissionen, 2013), and modern *B. napus* cultivars are down to 8-15 µmol/g, from the 60–100 µmol/g of older cultivars (Wittkop, B., Snowdon, R. J. & Friedt, W. 2009, Jhingan *et al.* 2023). Sinalbin (Sb), originally isolated in the 1830s from white mustard seeds (*Sinapis alba*), is part of the aromatics, a small glucosinolate group derived from tyrosine (Harun *et al.* 2020), (Figure 4). *SOT16* is an enzyme responsible for catalyzing the final step between sinalbin and its precursor 4-Hydroxybenzyldesulphoglucosinolate (Jain *et al.* 1989, Klein & Papenbrock, 2009). Limiting the conversion rate is theorized to potentially lower GL content in seeds. This gene, however, also acts as a catalyzing agent for multiple amino acid GL pathways (Harun *et al.* 2020). GL biosynthesis enzyme *BUS1* acts upstream of *SOT16* in the methionine-derived aliphatic GL pathway (Reintanz *et al.* 2001).

While not as well studied as the oil-related transcription factors (TFs), evidence suggests that *MYB28* acts as a positive transcriptional regulator for aliphatic GL (Gigolashvili *et al.* 2007). Its critical role in GL synthesis is further supported by a recent study in *B. napus*, showcasing a sharp reduction in aliphatic GL in a functional mutant of *MYB28* (Jhingan, 2023). Due to their involvement in GL-synthesis and prior identification in *L. campestre*, both *BUS1* and *MYB28* are of interest to improve seed cake nutritional value.

Aliphatic and indolic GL is proposed to be transported out of the cell by *GTR1* and *GTR2* (Harun *et al.* 2020). *GTR1* and *GTR2* are implicated in tissue-specific GL distribution and transport within the plant in *A. thaliana* (Andersen & Halkier, 2014, Andersen *et al.* 2013). Recent studies in *Brassica juncea* have shown that targeting *GTR1* and *GTR2* for silencing decreases GL accumulation in seeds (sink-tissue), while maintaining it in the leaves (source-tissue) (Nambiar *et al.* 2021, Kumari *et al.* 2022). As already stated, *GTR1* and *GTR2* have also been functionally validated in *L. campestre* (Sandgrind *et al.* 2022). This makes *GTR1* and *GTR2* interesting for improving the nutritional quality of the seed cake with minimal sacrifice of defensive properties.

The β-thioglucosidaes (myrosinases) enzymes are responsible for the degradation of the biologically inactive glucosinolates. Before activation, they are contained in specialized compartments, unable to interact with the GL (Rask *et al.* 2000). Upon release, they catalyze the hydrolyzation of GL into the toxic isothiocyanates (Tripathi & Mishra 2007, Lee, Kim & Woyengo, 2020). *PEN2*, commonly referred to as an atypical myrosinase, is a glycosyl hydrolase localized in the peroxisome,

associated with the pathogen-induced catalysation of cleavage of indole GL to its breakdown products (Bednarek *et al.*, 2009, Lipka *et al.*, 2005) (Figure 4). *PEN2* is of interest for decreasing the toxicity of *L. campestre* seed cake, due to its potential to decrease enzymatic cleavage and prevent subsequent toxic breakdown products from forming. *PEN2* has otherwise most notably been associated with the plant cell wall oomycete resistance (Lipka *et al.* 2005), microbe-associated molecular pattern (MAMP)-triggered callose formation due to bacterial infection (Bednarek *et al.* 2009) and hypersensitive response regulation (Johansson et al, 2014) in *A. thaliana*.

1.6 DNA sequencing methods

The earliest most well-known method for sequencing still in use today is Sanger-sequencing. Developed in 1970's, the method relies on easily detectable nucleotides specialized to terminate a Polymerase Chain Reaction (PCR) when incorporated. After termination, each product length can be read as a specific base-pair (Shendure *et al.* 2017). Multiple sequenced fragments can be overlapped in what is called a scaffold, to create larger continuous stretches of DNA. Sanger sequencing was revolutionary at this time, paving the way for large-scale sequencing, such as the Human Genome Project (Heather & Chain, 2016). Nowadays Sanger sequencing has partly been replaced by more high-throughput next-generation sequencing (NGS) methods such as short-read (50-300 bp) Illumina sequencing. Sanger sequencing is however still actively used for many applications including small-scale and validation studies, praised for its low error rate and long (>500 bp) sequencing reads (Heather & Chain, 2016).

The most widely used NGS method to date is Illumina sequencing, relying on nucleotides with fluorescent dyes emitting detectable signals when binding to the DNA template with the help of DNA polymerase (Heather & Chain, 2016). Illumina has the capacity for high throughput sequencing, with the downside of shorter reads, which may cause issues in assembling regions high in repeats. The method uses DNA-polymerase for strand synthesis, that have a small albeit existing error rate in replication that is crucial for evolution as it over time introduces genetic variation in natural populations (Lee *et al.* 2016). The use of DNA-polymerase in Illumina therefore causes a higher error rate than Sanger, with approximately 0.1-0.6% up to 15% failure, compared to 0.01% (Stoler & Nekrutenko 2021, Cheng & Xiao 2022, Liu *et al.* 2012, Shendure *et al.* 2008).

Sanger and NGS sequencing are suitable for different goals. Where Sanger is useful for initial discovery and validation, NGS is better suited for larger statistical analyses, whole gene sequencing and genome-wide association studies (GWAS) requiring high throughput. It should also be considered that not all target

organisms have the prior genomic information required for NGS and, researchers may therefore opt for Sanger sequencing out of necessity.

1.7 Genetic polymorphisms

Genetic polymorphism describes variation in the DNA sequence of individuals, groups, and species. For example, this variation can be insertion or deletion (INDEL) of bases, inversion of a sequence, or base substitution at a single locus, which is also called single nucleotide polymorphism (SNP).

SNPs (Figure 5) are found in both coding and noncoding regions. For example, SNPs were present on average every 124-131 bp in coding, and every 31-48 bp in non-coding regions in maize populations (Ching *et al.* 2002, Ghosh *et al.* 2002). SNPs found in coding regions can be either synonymous or non-synonymous. Synonymous mutations are silent in the sense that the amino acid remains unchanged. Any change in the amino acid sequence due to a mutation is known as a nonsynonymous mutation, which can be of missense or nonsense types. Missense mutations can affect the protein folding or protein interactions and consequently act negatively on function. The most severe mutations caused by SNPs is a nonsense mutation that results in a premature stop codon, which results in a truncated, possibly entirely non-functional protein.

Most cases of phenotypic variation are caused by missense mutations (Cubillos, Coustham & Loudet, 2012). SNPs correlated to beneficial traits have, however, also been found in non-coding regions. This is due to linkage disequilibrium between loci inside and outside the coding sequence (CDS) (Kruklyak, 1999), functional motifs in the promotor region affecting binding and mRNA-stability (Yvert, Brem & Whittle, 2003) as well as disruption of alternative splicing sites (Yang, Kim & Bhak, 2009). In fact, even synonymous mutations, which are traditionally considered neutral, may affect phenotype through for example altered mRNA, introduction of a promotor region (Lebeuf-Taylor, 2019) or codon bias influencing translation speed (Chu & Wei, 2021). Synonymous mutations have been linked to both increased and decreased fitness in *Pseudomonas* bacteria (Lebeuf-Taylor, 2019) and stronger selection in cancer-related genes in humans (Chu & Wei, 2019). There is however little research into the effects of synonymous mutations in plants.

SNPs are a great tool in molecular breeding, and they are used extensively for genome-wide association studies (GWAS), genetic diversity analysis of populations and MAS (Kumar, Banks & Cloutier 2012). Thus, SNPs can be very useful as markers when statistically linked to a phenotype. Due to the rapid advancement of next-generation sequencing and genome annotations in recent decades, SNP discovery is a relatively easy and cheap way to find beneficial genomic variation. sequencing projects, such as Restriction-site Associated DNA

(RAD) sequencing (Gustafsson, 2018), Genotyping-By-Sequencing (GBS) (Hammenhag *et al.* 2020, Geleta et al, 2020), and the low throughput Sanger sequencing (Gustafsson, 2018) have led to the discovery of thousands of SNPs that can be used for developing genomics-led breeding tools.

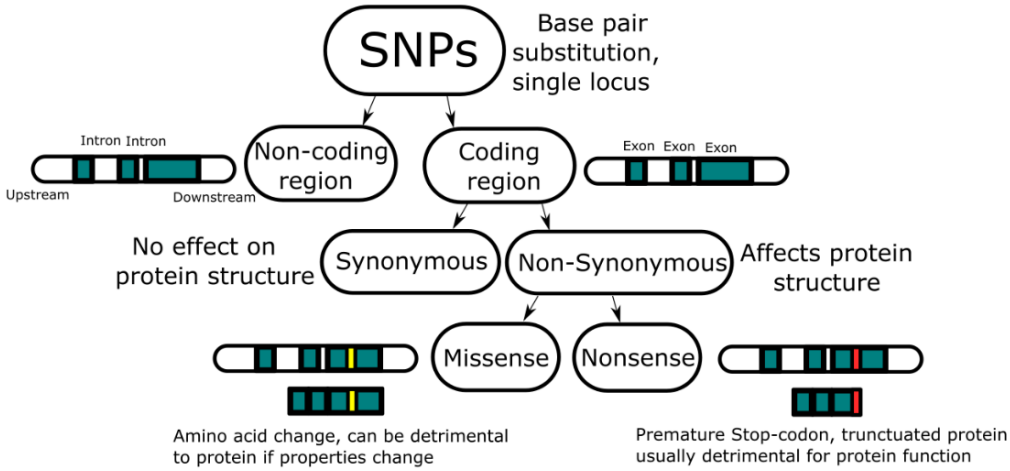


Figure 5. Schematic illustration of SNP classification and predicted impact on protein structure and function.

Other widely used markers in plant breeding to evaluate genetic diversity includes Restriction Fragment Length Polymorphism (RFLP), Random Amplified Polymorphic DNA (RAPD), Amplified Fragment Length Polymorphism (AFLP) and Single Sequence Repeats (SSR) (Hasan *et al.* 2021). RFLP and AFLP both relies on enzymatic cleavage sites which will vary between individuals, and after cleavage with enzymes, the lengths of sequences produced after cleavage can be analysed either directly or in the case of AFLP after a round of amplification (Hasan *et al.* 2021). RADP and SSR markers on the other hand are similarly to SNPs PCR-based. In RAPD the amplified fragments are based on primers binding to random unknown positions in the genome, which may be useful in species with limited genomic information (Hasan *et al.* 2021). SSR markers are short repeated sequences, where the number of times they are repeated varies between individuals. The regions flanking the repeats can be used as target for primers (Mason, 2015).

In theory, a polymorphism with a detrimental effect on protein structure and function will change the observed and measured phenotype of the individual in question. However, this is not always the case. Multiple genes may have a redundancy in their function, and when the expression of one decrease, another can instead perform its function (Nowak *et al.* 1997). Another scenario is that a given gene may interact with other genes, which may affect phenotypes in an unpredictable way. It is therefore important to keep in mind that functional redundancy and gene networks of key genes may cause hindrance to the discovery of novel beneficial markers.

1.8 Promoting *L. campestre* domestication through interspecific hybridization

A key challenge for traditional plant breeding is how to handle limited natural variation and a lack of beneficial traits within populations of interest. A genetic bottleneck generally occurs over time in plant domestication due to extensive selective pressure for a few traits in a limited number of populations (Meyer & Purugganan, 2013). Low genetic diversity is a concern in *L. campestre* breeding (Gustafsson *et al.* 2018). An example of this is its naturally highly shattering pods, where no resistant lines in *L. campestre* have been identified. One way forward can be to evaluate closely related species for beneficial traits and diversity, which may be introduced via interspecific hybridization. Such attempts have successfully been made in *L. campestre* in the past to expand genetic diversity, by crossbreeding it with *L. heterophyllum* and *L. hirtum* (Gustafsson *et al.* 2018). The interspecific hybridization of *L. campestre* with *L. heterophyllum* has resulted in some derived lines with high resistance to pod shattering, which proves that this can be an effective approach to improve key traits (unpublished, Mulatu Geleta).

Additional interesting candidate species that are superior to *L. campestre* in certain traits have been identified. These are *L. graminifolium* (Nilsson, Johansson & Merker 1998) for oil quality and content and *L. draba* for pod shatter resistance (Geleta *et al.* 2013). Neither of these species has yet been successfully hybridized with *L. campestre*.

In his study, 12 relatives of *L. campestre* were evaluated for the purpose of future interspecific hybridization (Table 2).

Breeding barriers can easily occur within a genus when the species are not phylogenetically closely related, when they differ in chromosome number, ploidy level or due to other factors – resulting in embryo abortion when hybridized (Kopecký, Martín, & Smýkal, 2022). Embryo rescue protocol, which can be used to overcome embryo abortion due to incompatibility, has been successfully established for other Brassicaceae species (Ripa *et al.* 2020, Wen *et al.* 2008) but not yet for *Lepidium.* Minor studies have, however, been performed and some advances have been made but more work is required (Reyes Esteves 2021, Fahlgren 2014). Now, this limits the current candidates for successful interspecific hybridization. Despite the potential in using close relatives to enhance *L. campestre* domestication and development, few studies have been performed to evaluate *Lepidium* species for use in interspecific hybridization with *L. campestre*.

Table 2. Morphology, known traits, and general habitat for *Lepidium* species evaluated in this study for future interspecific hybridization with *L. campestre*. Species with multiple reported chromosome numbers are described as 2n = X, Y.

[a]*Warwick & Al-Shehbaz (2006),* [b]*Nilsson, Johansson & Merker (1998),* [c]*Bona (2014),* [d] *Kew (2023b),* [e]*Escudero et al. (2000).*

Species	Ploidy [a, c]	Common name	Perenniality	Morphology	Other traits	Habitat	Geography
Lepidium campestre	2n = 16	Field cress	Annual or Biennial [c]	15–60 cm, upright [c]	Cold hardy [b,f]	Field margin, roadside. forests [c]	Europe, naturalized Australia and North and South America [a]
Lepidium cardamines	2n = 16	No common English name	Biennial [d]	70 cm, upright	Tiny seeds	temperate biome [d]	Spain [a]
Lepidium graminifolium	2n = 16, 48	Grassleaf pepperweed	Perennial [c]	10–80 cm [c]	No close relatives [b]	Dry slopes, s.l.–1800 m. [c]	Europe, Northwest Africa, Southwest Asia [c], S. Central Europe to Medit. [d]
Lepidium heterophyllum	2n = 16	Smith's pepperworth	Perennial [c]	Upright [d]	Not cold hardy [b]	Temperate biome [d]	Algeria, France, Great Britain, Ireland, Morocco, Portugal, Spain [a]
Lepidium hirtum	2n = 16	Mediterranean pepperweed (general)	Perennial [b]	up to 17 cm, not upright [b]	Morphologically unfavourable [b]	Forests, fields, roads [c]	Algeria, Corse, East Aegean Is., France, Greece, Kriti, Lebanon-Syria, Morocco, Sicilia, Spain [a]
Lepidium perfoliatum	2n = 16	Clasping pepperweed, Various leaved pepperwort	Annual or biennial [c]	10–35 cm, upright [c]	Invasive	Waste places, roadsides, salty grounds, stony, slopes, s.l. 1850 m [c]	Central and South Europe, Northwest and Central Asia [a]
Lepidium ruderale	2n = 16, 32	Roadside pepperweed, Narrow leaved pepperweed	Annual [c]	10–30cm, upright [c]		Roadsides, waste places, field margins, s.l.–1550 m. [c]	South and West Australia [a], Europe [d]
Lepidium sativum	2n = 16, 24	Garden cress, Common cress	Annual [c]	10–70 cm, upright [c]	Fast-growing	waste moist places [c]	Semi-domesticated, Cultivated all over the world [d]
Lepidium spinosum	2n = 16, 24	No common English name	Annual [c]	10–60 cm, upright [c]	Apetalous , invasive	Moist slopes, meadows, field margins, s.l.–1300 m [c]	West Syria [a]
Lepidium subulatum	2n = 16, 32	No common English name	*Uknown*	Upright [d]	Gypsophite (= thrives in Hydrated Calcium Sulphate)[e] Tiny seeds	Arid temperate regions[d]	Algeria, Morocco, Spain [a]
Lepidium vesicarium	2n = 16, 32	Bladdery pepperwort	Annual or Biennial [c]	15–70 cm, upright [c]	nodal swelling, no close relatives [b]	Stony slopes, roadsides, 800–1900 m [c]	Caucasus, Iran [a]
Lepidium virginicum	2n = 16, 32	Virginian Pepperwort	Annual or Biennial [c]	10–70 cm, upright [c]	Invasive	Roadsides [d]	Europe, North America, Australia [a]

2. Material & Method

2.1 Plant material

Seeds from 40 accessions of *Lepidium campestre* (Appendix 1), with previously measured oil content, oil composition, and glucosinolate content were sown in a greenhouse at SLU, Alnarp. The soil used for planting was *Emmaljunga Exklusiv Blom och Plantjord*. Five seeds or seedlings of each accession were planted in a 3 L plastic pot filled with soil. Basal leaves of five-week-old plants were sampled for DNA and RNA extraction in 2 ml microcentrifuge tubes containing two glass beads, flash frozen, and stored in liquid nitrogen until DNA and RNA extraction.

For characterizing relatives of *L. campestre*, 32 accessions of 11 *Lepidium* species (Appendix 2) were acquired through the gene banks: United States Department of Agriculture, Agricultural Research Service (USDA-ARS), and IPK Gatersleben (IPK). These species are *L. cardamines (*1 acc.), *L. graminifolium* (1 acc.), *L. heterophyllum* (1 acc.), *L. hirtum* (4 acc.), *L. perfoliatum* (3 acc.), *L. ruderale* (12 acc.), *L. spinosum* (2 acc.), *L. subulatum* (1 acc.), *L. sativum* (1 acc.), *L. vesicarium* (1 acc.) and *L. virginicum* (5 acc.). Planting and sampling for DNA and RNA extraction of these accessions are as described for *L. campestre.*

2.2 Bioinformatic analysis

2.2.1 Gene structure predictions in *L. campestre*

Exon/intron structure was predicted for all 15 studied genes using BLAST sequence alignment with *L. campestre* genomic sequences and *A. thaliana* Coding Sequence (CDS) from The *Arabidopsis* Information Resource (TAIR). Predicted splice sites were further validated using prediction from NetGene2 web software version 2.4. CDS sequences of all examined genes were translated to predicted protein using the Expansy online tool, to validate if prediction would result in a functional protein. This was successfully done for all genes except for *WRI1*, where a general intron/exon structure could be suggested but no stop codon could be predicted.

PEN2 was further validated with experimental mRNA PCR sequencing data (section 2.4).

2.2.2 Bioinformatic study of Brassicaceae species

An in-depth bioinformatics study on *FAE1_1*, *FAD2*, *PEN2* and *SOT16* was performed using the acquired genomic sequence of the respective genes of *L. campestre* whole genome sequence data. These gene sequences were used as baits in National Center for Biotechnology Information (NCBI) Basic Local Alignment tool (BLAST) to identify homologues in other Brassicaceae species and in NCBI Conserved Domains to find regions harbouring conserved domains. Previous bioinformatics studies were also used to find homologs for *FAD2* (Yang *et al.* 2012) and *FAE1* (Sun *et al.* 2013). Alignments were carried out using the MAFFT (with L-ins-I settings) algorithm included in Jalview version 2.11.2.6 (Waterhouse *et al.* 2009). Visualization of the aligned sequences was made based on BLAST results from homologous sequences graphic alignment, imported into Inkscape version 1.0.1 (2020).

2.3 Polymorphism discovery in *L. campestre* genes

In this study, 15 *L. campestre* genes related to oil content, oil composition, and glucosinolate biosynthesis, degradation and transport were evaluated for polymorphisms (Table 3).

Table 3. List of genes used for genetic variation analysis in this study. Coding sequence predicted from *L. campestre* genomic sequence after alignment with *A. thaliana* CDS. Full name and function based on data gathered from The *Arabidopsis* Information Resource (TAIR, 2023) based on functional studies in *A. thaliana*. CDS = Coding sequence, VLCFAs = Very long chain fatty acids.

Gene	CDS length (bp)	Full name	Related to	Associated Function
FAD2	1152	*FATTY ACID DESATURASE 2*	Oil composition	(18:1) to (18:2) fatty acid conversion
FAE1_1	1521	*FATTY ACID ELONGASE-1_1*	Oil composition	Biosynthesis of VLCFAs
FAE1_2	1467	*FATTY ACID ELONGASE-1_2*	Oil composition	Biosynthesis of VLCFAs
FAE1_3	1460	*FATTY ACID ELONGASE-1_3*	Oil composition	Biosynthesis of VLCFAs
KCS8	1446	*3-KETOACYL-COA-SYNTHASE-8*	Oil composition	Biosynthesis of VLCFAs
LEC1	1679	*LEAFY COTYLODON-1*	Oil content	Fatty acid biosynthesis
LEC2	3530	*LEAFY COTYLODON-2*	Oil content	Fatty acid biosynthesis
TAG1	3102	*TRIAGLYCEROL-1*	Oil content	Fatty acid biosynthesis
WRI1	Not yet determined	*WRINKLED-1*	Oil content	Fatty acid biosynthesis
BUS1	1614	*BUSHY-1*	Glucosinolate content	Glucosinolate biosynthesis
SOT16	1014	*SULFOTRANSFERASE-16*	Glucosinolate content	Glucosinolate biosynthesis

MYB28	1299	*MYB DOMAIN PROTEIN-28*	Glucosinolate content	Glucosinolate biosynthesis
PEN2	2498	*PENETRATION-2*	Glucosinolate content	Glucosinolate hydrolyzation (into breakdown products)
GTR1	2571	*GLUCOSINOLATE TRANSPORTER-1*	Glucosinolate content	Glucosinolate transport
GTR2	2440	*GLUCOSINOLATE TRANSPORTER-2*	Glucosinolate content	Glucosinolate transport

2.3.1 DNA Extraction

Genomic DNA was extracted according to Gustafsson *et al.* 2018 from harvested leaf tissue that was flash-frozen in liquid nitrogen and homogenized by shaking with glass beads in a Mixer Mill MM400 (Retsch, Haan, Germany) for 30 x 2 seconds, at 30 Hz. After homogenisation, the samples were incubated for 1 hour at 52°C in preheated CTAB-buffer (0.1 M Tris, 20 mM EDTA, 1.4 M NaCl, 2% CTAB, pH 7.5). Samples were centrifuged for 15 minutes at 13,200 rpm in an miniSpin (Eppendorf , Hamburg, Germany). From each sample, 400 µL of supernatant was transferred to a new microcentrifuge tube and 400 µL 24:1 chloroform/isoamylalcohol was added. Samples were then centrifuged for 3 minutes at 12, 000 rpm and 350 µL of the top phase was transferred to a new 1.5 mL tube with 350 µL ice-cold isopropanol and centrifuged for 3 minutes at 12 ,000 rpm. The supernatant was discarded. The obtained DNA pellet was washed using 350 µL wash buffer (7,6 ml 100% ethanol, 33 µL 3M NH_4Ac, MilliQ up to 10 ml), and centrifuged for 3 minutes at 12, 000 rpm, before discarding the supernatant. The last drops of wash buffer were removed by pipette after centrifuging for 1 min at 12,000 rpm. Finally, the samples were dried and resuspended in a 50 µL TE-buffer (0.1 ml TrisHCl 1M pH 8, 20 µL EDTA 0.5 M pH 8, MilliQ up to 10 ml). The quantity and quality of the extracted DNA were checked by running it on a 1.2% agarose gel and measured with a NanoDrop spectrophotometer DS-11 FX (DeNovix, Wilmington, USA).

2.3.2 PCR and Primer design

Gene sequences from *L. campestre* were retrieved from unpublished whole genome sequence data through comparative genomics analysis using *A. thaliana* gene sequences as baits.

The CDS of *L. campestre* genes were predicted by aligning *A. thaliana* CDS from TAIR with *L. campestre* genomic regions using BLAST search. Primers were designed to target coding regions using Primer3Plus, with advanced options set for a target region size of 500-1000 bp. Melting temperature (T_m) was calculated using the Phusion polymerase option with T_m Calculator (Thermo Scientific). Primers targeting the gene *FAE1* were carefully designed to result in amplification of *FAE1_1*, and not *FAE1_2* or *FAE1_3*, using MAFFT

(with default settings) alignment in Jalview software version 2.11.2.6 (Waterhouse *et al.* 2009), and mapping non-conserved regions between the three homologs. For a list of all primers designed, see Appendix 3.

All primers were tested for optimal melting temperature through PCR by determining the temperature that resulted in a single visible band of expected size, on an agarose gel. PCR was performed on 25 accessions selected as a diversity panel for the studied traits. The PCR program was x40 cycles lasting 1:05 h with denaturation at 98°C, annealing at 62°C, and primer extension at 72°C (Figure 6). All PCR products were run on 1.5% agarose gel with 0.5 µL GelRed/10 ml agarose for visualising the DNA using a BioDoc-It Imaging Systems camera (UVP). After confirming the bands on agarose gels, PCR products were purified with GeneJET PCR Purification kit (Thermo Scientific) following the manufacturer's protocol, except for the final elution step. Elution-columns in purification were incubated for 5 minutes at 70°C with elution buffer prior to final centrifugation. In cases of low concentration, this step was repeated twice with half of the desired elution volume each time. Finally, purified samples were measured for concentration and quality with a NanoDrop spectrophotometer DS-11 FX (DeNovix, Wilmington, USA). For Sanger sequencing, samples were sent to Eurofins (Cologne, Germany) after preparing them, following their "premixed clean PCR-product" option. For targeted next-generation sequencing, 9 to 47 µl of genomic DNA corresponding to 1500 ng, with OD-ratio of 260/280 and 230/260 around 1.85-2.0 and 1.4-2.0, respectively, were sent for sequencing to CD-genomics (New York, USA).

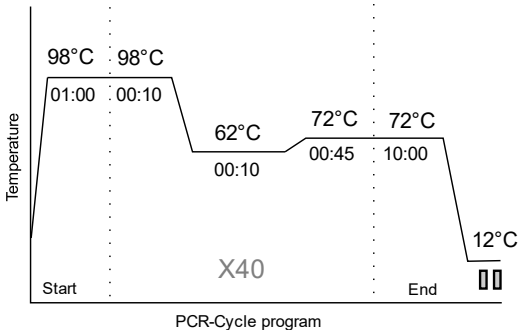


Figure 6. Phusion polymerase PCR program.

2.3.3 Sequencing

Two different sequencing approaches were used in this study. Sanger sequencing was performed on the gene regions of *FAE1_1*, *FAD2*, *PEN2* and *SOT16* for the diversity panel of 25 *L. campestre* accessions at Eurofins, utilizing both forward and reverse primers for sequencing with the option to record heterozygote sites. SNPs were identified by aligning the sequenced data with the target regions of the

reference genome with MAFFT (using default settings) algorithm included in Jalview software version 2.11.2.6 (Waterhouse *et al.* 2009).

Extracted genomic DNA samples of the 40 *L. campestre* accessions were sent to CD-genomics for NGS Illumina sequencing, and SNP discovery in 15 genes; *FAE1_1*, *FAE1_2*, *FAE1_3*, *FAD2*, *KCS8*, *LEC1*, *LEC2*, *TAG1*, *WRI1*, *PEN2*, *SOT16*, *MYB28*, *GTR1*, *GTR2* and *BUS1* (Table 3). In addition to the gene sequences, approximately 500-700 bp of the upstream and downstream regions of the target genes were analysed.

2.3.4 Statistical analysis on genetic variation

In R-studio, an unpaired t-test was performed between previously measured mean values in traits of interest including oleic acid (OA) content, erucic acid (EA) content, total glucosinolate content, Sinalbin (Sb) and Glucoallysin (Gla) content. Mean phenotypic values of the accessions with alternative allele were compared with that of accessions with reference allele at each locus. The statistical significance of the variation between the two groups of accessions was tested at 95% confidence level ($p < 0.05$). Only polymorphisms observed in two or more accessions were included in the statistical analysis.

Analysis of missense polymorphisms found in Sanger and NGS sequencing were performed using the web-based software Polyphen-2 to detect detrimental SNPs.

Multivariate analysis for SNP-based clustering was performed in JMP® software version 16.2 (2021) with multiple correspondence analysis to evaluate if sample phenotypes or geographical origins could explain observed variations between accessions.

2.4 RNA analysis of *PEN2*

2.4.1 RNA Extraction and cDNA synthesis

RNA was extracted from sampled and flash frozen *L. campestre* eight-week-old leaf tissue using RNeasy Plant Mini Kit (Qiagen, Germantown, USA). DNase treatment was performed using RNase free DNase I (Thermo Scientific, Vilnius, Lithuania) followed by cDNA synthesis performed using Revert Aid First Strand cDNA synthesis kit (Thermo Scientific, Vilnius, Lithuania). Primers used for cDNA synthesis were oligoDT, *PEN2*-mRNA-primer 5R, and *PEN2*-mRNA-primer 3R (Appendix 1).

2.4.2 PCR and Primer design for amplifying RNA

RNA primers for *PEN2* were designed to overlap a predicted exon-exon junction, see *PEN2*-bioinformatics (3.1.3) for exon prediction. Primers were used in PCR as described in (2.2.2. PCR and Primer design), see Appendix 3 for all primers.

PCR products were run on 1.5% agarose gel and purified with GeneJET Gel Extraction (Thermo Scientific, Vilnius, Lithuania) according to manufacturer's instructions. Purified PCR-products were measured using a NanoDrop spectrophotometer DS-11 FX (DeNovix, Wilmington, USA) and sent for sequencing to Eurofins after preparing them following their "premixed clean PCR-product" option. Sequenced products were aligned to *PEN2* genomic region with MAFFT (with default settings) algorithm using Jalview software version 2.11.2.6 (Waterhouse *et al.* 2009).

2.5 Characterization of *Lepidium* species

2.5.1 Analysis of agronomically important traits

Fifty seeds of each accession were grown in a Petri dish with x3 filter papers and Milli-Q water. Moldy seeds were immediately removed to prevent contamination of other seeds in the Petri dishes. Seeds were supplied with Milli-Q water every week or when the filter papers dried out. The number of germinated seedlings was counted on day 4, 8 and 12. Accessions with less than eight (<16%) germinated seeds after 12 days had 2x 25 new seeds placed on new petri dishes and were supplemented with 0.25% KNO_3 milliQ-H_2O and kept for seven days at 4°C. Germinated seedlings were transferred with a tweezer from Petri dishes to pots filled with soil. Up to ten seedlings of each accession were transplanted except for *L. hirtum subsp. atlanticum*, *L. hirtum subsp. nebrodese* and *L. hirtum* (PI 633256, PI 650270, LEP 81), for which only five seedlings were obtained, *L. virginicum subsp. menziesii* (Ames 31357), and *L. ruderale* (PI 633255) for which only one seedling was obtained and *L. virginicum* (Ames 34741, Ames 34742) for which no seedlings were obtained.

Seed traits

To analyse seed traits, 50-200 seeds from 34 *Lepidium* accessions including two *L. campestre* accessions, were counted using a Marwin seed analyser (Marvitech, Wittenburg, Germany) for seeds with width or length > 0.5 mm. Samples below the instrument's threshold were manually counted. The same instrument was used to determine the average weight, width and length of the seeds.

Flowering time and seed setting
Plants were monitored every other day for flower induction through petal opening, and seed setting through pod formation and loss of petals. At day-87, and again at day-110, all sufficiently grown biennial or perennial accessions were moved into a vernalization chamber (4°C) for eight weeks. Pots were fertilized with 4 g/L of fertilizer *Osmocote Exact Standard 5-6M* at the break of vernalization and moved back to the greenhouse.

The morphologies of seedlings (cotyledon leaf shape), basal leaves (young basal leaf shape and size), flowers (inflorescence shape, pedicel length, petal and sepal size and colour), pods (mature pod shape) and stem growth (upright/horizontal) were observed, photographed at multiple time points, and carefully sketched.

2.5.2 *Lepidium* seed oil extraction

Oil extraction of 16 *Lepidium* accessions was performed by weighing ten seeds of each accession, in three replicates. Seeds were placed in a wide glass tube, and 1 ml of 0.15 M HAc and 3.75 mL $CHCl_3$/MeOH were added and the seeds were homogenized with an Ultraturrax rod 5x30 seconds. The rod was cleaned with ethanol three times between the homogenization of each sample. Thereafter, 1.25 ml $CHCl_3$ and 0.9 mL milli-Q water were added and the samples were vortexed for 30 sec. Finally, the samples were transferred to a glass tube with screw lids.

For seeds that weighed below 0.1 mg/seed (*L. cardamines & L. subulatum*), the seeds were crushed manually in a large tube with a glass rod, to which the HAc and $CHCl_3$/MeOH were added before they were transferred into a glass tube with a screw lid. Additional $CHCl_3$ and milli-Q were added to the large tube to dissolve remaining residues and poured into glass tubes. The sample solutions in the glass tubes were vortexed for 30 sec.

The samples were centrifuged for 2 min at 3000 rpm. The lower $CHCl_3$ phase was transferred to a glass tube with screw a lid. From the glass tube, 200 µL was transferred to a methylation-safe glass tube with a screw lid and left to dry for a few minutes under a beam of nitrogen. Once the samples were dry, a 17:0-Me standard solution and 2 mL methylation solution (2% H_2SO_4 in methanol) were added to the samples. Then, the lids of the sample tubes were tightly closed and left to methylate at 95°C for 45 minutes using a heating block. The samples were then removed from the heating block and allowed to cool down, before 1 mL milli-Q water and 0.75 mL heptane were added and the samples were centrifuged for 2 min at 2000 rpm. Two hundred µL of the upper heptane phase was transferred to a GC-vial. Then, 2 µL of this solution was injected into a Gas chromatograph (GC) (Agilent, model 7890A), equipped with a wax column (WCOT Fused Silica CP-Wax 58) and FID detector (Agilent, Santa Clara, CA, USA).

2.5.3 Evaluating the effect of selected polymorphisms in *FAD2*, *LEC2*, *WRI1* and *TAG1* in *Lepidium* species

Fourteen interesting polymorphic sites with a significant correlation with seed oil content were selected to be further evaluated in 14 *Lepidium* species (Appendix 2). This was not done for *L. cardamines*, as its seedlings were not obtained.

First, the homologous sequences of the genes of interest, including their 500 bp upstream and downstream regions, in *A. thaliana*, *A. arenosa*, *B. napus* and *B. rapa* were found by using *L. campestre* as query bait in NCBI BLAST. Primers were then designed for highly conserved regions, detected by MAFFT alignment (with L-ins-I settings) in Jalview version 2.11.2.6 (Waterhouse *et al.* 2009). Primers were designed using Primer3Plus and T_m was estimated by the T_m calculator (Thermo Scientific).

The DNA samples from the 14 *Lepidium* species were extracted from leaf tissue as described above (see section 2.2.1, DNA Extraction). PCR was performed as described above (see section, 2.2.2. PCR and Primer design), purified with GeneJET PCR Purification kit (Thermo Scientific, Vilnius, Lithuania) according to manufacturer's instruction, and sent for sequencing to Eurofins as described above (see section 2.2.3, Sequencing).

The sequences of the PCR products were then aligned using the MAFFT alignment tool (with L-ins-I settings) in Jalview version 2.11.2.6 (Waterhouse *et al.* 2009) to detect the level of their conservation. Sequence similarity was determined using NCBI BLAST multiple sequence alignment tool, with mega blast setting using *L. campestre* sequences as a query. Visualization of sequence conservation was made based on BLAST results from multiple sequence alignment with E-value between 0-1.00E^{-40}. Sequences were downloaded as SVG files, with simplified colouring and imported into Inkscape version 1.0.1 (2020) for further image processing.

A neighbour-joining phylogenetic tree was constructed using Jalview version 2.11.2.6 (Waterhouse *et al.* 2009), based on the longest continuous stretch of the successfully sequenced region of the *FAD2* gene in all studied accessions. *FAD2* was chosen, as it is the most conserved among the studied genes.

2.6 Sequencing method analysis

A cost-effectiveness calculation was performed between the NGS and Sanger sequencing. Comparing the cost for NGS sequencing of 15 genes including 700

bp before the start codon (ATG) and after stop codon (CD-genomics), with the theoretical cost of performing the same work with Sanger sequencing (Eurofins).

For calculation of costs of the Sanger sequencing, the following costs were included: Phusion polymerase (Phusion™ High-Fidelity DNA Polymerase, F530S, Thermo Fisher, Vilnius, Lithuania), dNTP, R0191, Thermo Scientific, Vilnius, Lithuania) and primer pairs (oligo primer, Sigma Aldrich) based on a rough calculation of 1 primer pair per 650-800 bp sequenced and PCR-purification (GeneJET PCR Purification Kit, K0701, Thermo Scientific, Vilnius, Lithuania), pre-paid labels for tubes (TubeSeq Labels, 3094-000PPB, Eurofins, Köln, Germany) or pre-paid plates for pre-purified (PlateSeq Kit, Eurofins, Köln, Germany) or unpurified (PlateSeq Kit Mix, Eurofins, Köln, Germany) PCR-products.

The cost associated with NGS sequencing included primer setup and sequencing of 15 genes, where a reference genome for the regions was provided (Targeted region sequencing, CD-genomics, New York, USA). Additional costs for NGS included shipping costs of pre-extracted genomic DNA.

3. Results

3.1 Bioinformatic analysis of *FAD2*, *FAE1*, *PEN2* & *SOT16* in selected Brassicaceae species

Specific accession numbers and sequence lengths can be found in Appendix 5.

3.1.1 *FAD2*

The gene *FAD2* has previously been shown to be associated with total seed oil content in *L. campestre* and relatives (Ivarsson *et al.* 2016, Sandgrind *et al.* 2023). In purpose of studying the gene structure of *FAD2* in species of Brassicaceae with available genome annotations, gene sequences from close relatives of *L. campestre* were retrieved from NCBI and used for comparative analysis. *FAD2* is present with multiple homologous copies in the genome of *C. sativa* (3), *B. napus* (4), *B. rapa* (2) and *B. oleracea* (2) but there is only a single copy of *FAD2* in *L. campestre*. The comparative analysis showed that the level of conservation for the gene structure *FAD2* is high, with few INDELs and high sequence similarity (Figure 7). The sequence similarity ranges from 91 % (*A. thaliana*) to 82% (*B. oleracea* C01). Notably however is the large deletion present at bp position 159-165 and highly polymorphic and deletion-rich region in the following 100 bp in the pseudogenes from *B. napus* and *B. rapa*, which potentially could be partly responsible for rendering them non-functional, as these features are not shared with other *FAD2* homologues. A species-specific 3-6 bp long insertion can be found in all gene copies in *C. sativa* at bp position 29. Another genus specific 3 bp insertion can be found across all *Brassica* species at 610 bp. No INDELs are present within the conserved motifs.

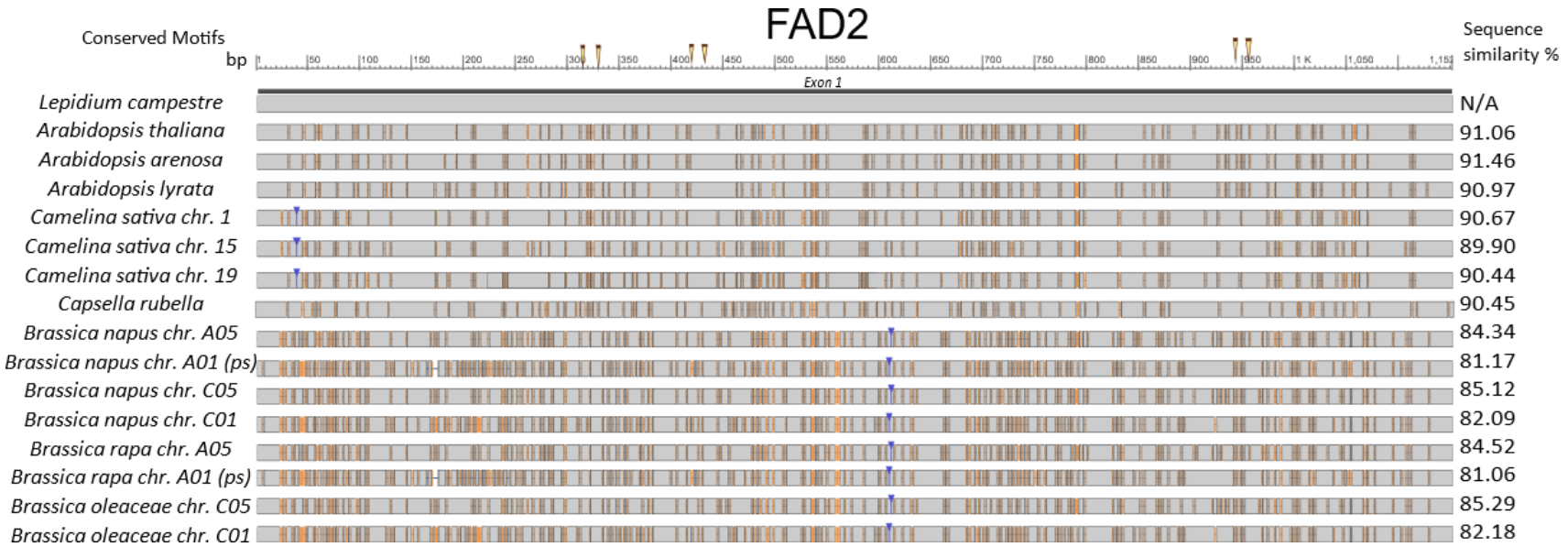


Figure 7. Visualization of *FAD2* gene conservation and sequence similarity across selected Brassicaceae species compared to *L. campestre*. Orange indicates a polymorphism, blue indicate an insertion. Regions of conserved histidine cluster motifs are marked with yellow triangles. ps – pseudogene, chr. – Chromosome

3.1.2 *FAE1*

The gene *FAE1* has previously been shown to be associated with seed oil composition, more specifically VLCFA and erucic acid content. It is well studied in *B. napus* (Wang *et al.* 2022, Stefansson & Hougen, 1964) but has also been validated in *L. campestre* (Ivarsson *et al.* 2016). For studying the gene structure of *FAE1* in species of Brassicaceae with available genome annotations, gene sequences from close relatives of *L. campestre* were retrieved from NCBI and used for comparative analysis. *FAE1* is present with multiple homologous copies in the genome of *L. campestre* (3), *C. sativa* (3), *B. napus* (2) and as a single copy in *A. thaliana, A. lyrata, C. rubella,* and *T. arvense*. The conservation level for FAE appears to be quite well conserved, with few INDELs (Figure 8). Generally, the first 10 bases appear non-conserved across most species. A conserved 3 bp insertion of ATA can be found across the *Brassica* species at 135 bp, and in the same locus a 3 bp insertion of CTA in *T. arvense* (Figure 8).

Retrieved sequences for genes Lcamp*FAE1_2* and Lcamp*FAE1_3*, identified as *FAE1*-homologues *L. campestre*, share low sequence similarity to Lcamp*FAE1_1* (73.3%, 75.9%). They are therefore excluded from the bioinformatic analysis.

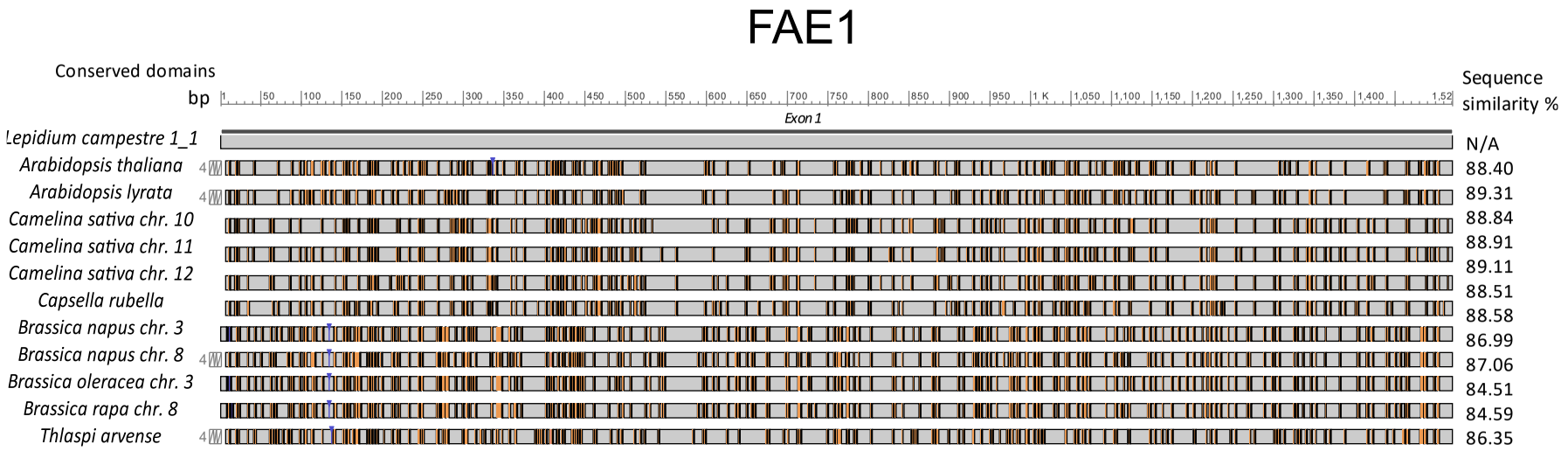


Figure 8. Visualization of *FAE1* gene conservation and sequence similarity across selected Brassicaceae species compared to *L. campestre*. Orange indicates a polymorphism, blue indicate an insertion.

3.1.3 *PEN2*

The gene *PEN2* has previously been shown to be associated with GL hydrolysis in *A. thaliana* (Bednarek *et al.*, 2009, Lipka *et al.*, 2005).

The *PEN2* exon/intron structure could not be established using only alignment between *A. thaliana* CDS and *L. campestre* genomic information in the same way as for the other genes in this study due to the highly polymorphic sequence and complex intron/exon structure of *PEN2* (Figure 9A). A project was therefore undertaken to establish the intron/exon structure experimentally using *A. thaliana* CDS as a basis for a prediction. From sequencing of amplified *PEN2* mRNA based on primers in Figure 9, it was concluded that *PEN2* has an exon/intron structure similar to *A. thaliana* with 10 exons and the presence of a very small first exon of only 10 bp. The only exon-exon junction not fully validated is between exon four and five, for which a band of correct length was obtained on gel. The PCR-product was however lost after being sent to the sequencing facility.

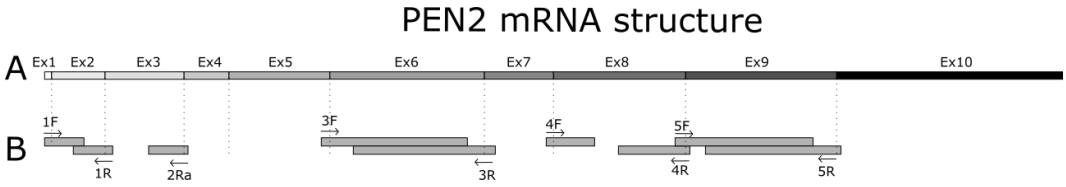


Figure 9. PCR-primer products for mRNA-characterization (B) aligned to the predicted mRNA product of *L. campestre PEN2* (A). Prediction based on *A. thaliana PEN2* structure. A PCR-product was obtained and verified on gel for the region spanning exon four to five, but not sequenced due to package handling issues, and hence not shown. Abbreviations: Ex = exon, F = forward primer, R = reverse primer.

To study the gene structure of *PEN2* in species of Brassicaceae with available genome annotations, gene sequences from close relatives of *L. campestre* were retrieved from NCBI and used for comparative analysis. *A. alpina* was included due to its sequence annotation on the antisense strand compared to *L. campestre* and *B. napus.* It was evaluated whether there was the presence of an antisense transcript in *L. campestre* as well, supported by the presence of multiple splice sites with high credibility predictions by NetGene2 web software version 2.42. A test was performed with one set of primer pairs; it did however not yield any product suggesting that there is no antisense transcript.

PEN2 is present with one homologous copy in all studied Brassicaceae species (Figure 10). There is a relatively low level of conservation for *PEN2*, especially in the introns (Figure 10). The exons are however also highly

polymorphic – with an insertion found in both exon 9 and 10 across at least some of the studied species. An insertion in the middle of exon 9 at 1931 bp is conserved across all species except for *L. campestre*. The insertion is longer in *A. thaliana* (104 bp) and *A. arenosa* (95 bp) than in the *Brassica*s (87 bp). The insertion in exon 10, conserved across the *Brassica* species, includes two closely followed A/T and repeat-rich segments of 132 and 120 bp. Three highly polymorphic regions can be identified where there are insertions and deletion present in the same loci in different species (Figure 10). In the first intron in bp position 68, one such highly polymorphic region site is present in the same area where the studied *Arabidopsis* species have an insertion, while the *Brassica* have a deletion. Another similar site is also present in intron 7 at bp position 1220 and intron 9 at 2091 (Figure 10). Two INDELs inside exon 10 appear to be specific to the *Brassica* species, with a 117-120 bp insertion at position 2102 and a 27 bp deletion at position 2377 (Figure 10).

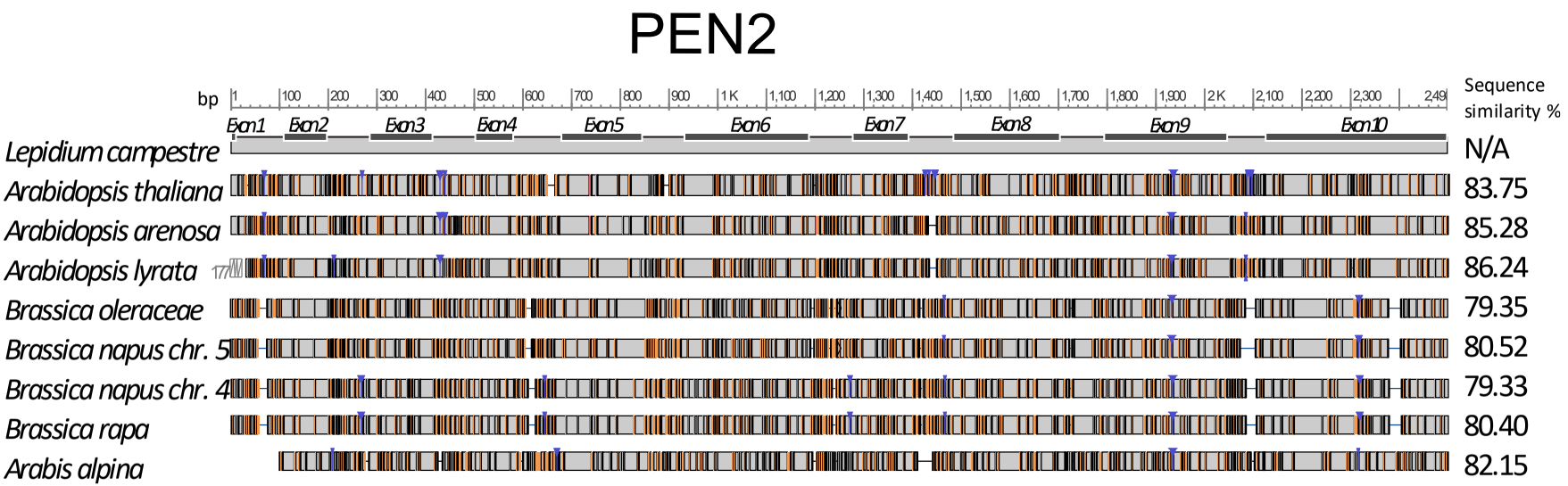


Figure 10. Visualization of *PEN2*-conservation across selected Brassicaceae species compared to *L. campestre*. *L. campestre PEN2*-Exon (E)/ Intron (I) gene structure based on CDS-sequences from *A. thaliana* (TAIR, 2023) supported by experimental data (Figure 9). Orange indicates a polymorphism, and blue indicates an insertion.

3.1.4 *SOT16*

The gene *SOT16* has previously been shown to be associated with GL synthesis in *A. thaliana* and is of interest due to its association with *L. campestre'*s major GL sinalbin (Jain *et al.* 1989, Klein & Papenbrock, 2009, Harun *et al.* 2020). To study the gene structure of *SOT16* in species of Brassicaceae with available genome annotations, gene sequences from close relatives of *L. campestre* were retrieved from NCBI and used for comparative analysis. *S. alba* was included because it is among the few Brassicaceae with sinalbin as a major GL (Harun *et al.* 2020).

SOT16 is present with multiple homologs in *C. sativa* (3), *B. napus* (2) and *S. alba* (2). Generally, the overall gene structure of *SOT16* appears to be conserved, however with smaller non-conserved regions (Figure 11). Notably all *Brassica* species share a 6 bp insertion around position 31 as well as a 3 bp insertion at 556 bp. An insertion of 3 bp is present in the *Arabidopsis* species at position 34 bp. It is possible that the chromosome 16 *SOT16* in *C. sativa* is not actually a homolog but rather a similar sequence or a pseudogene due to its shorter sequence length.

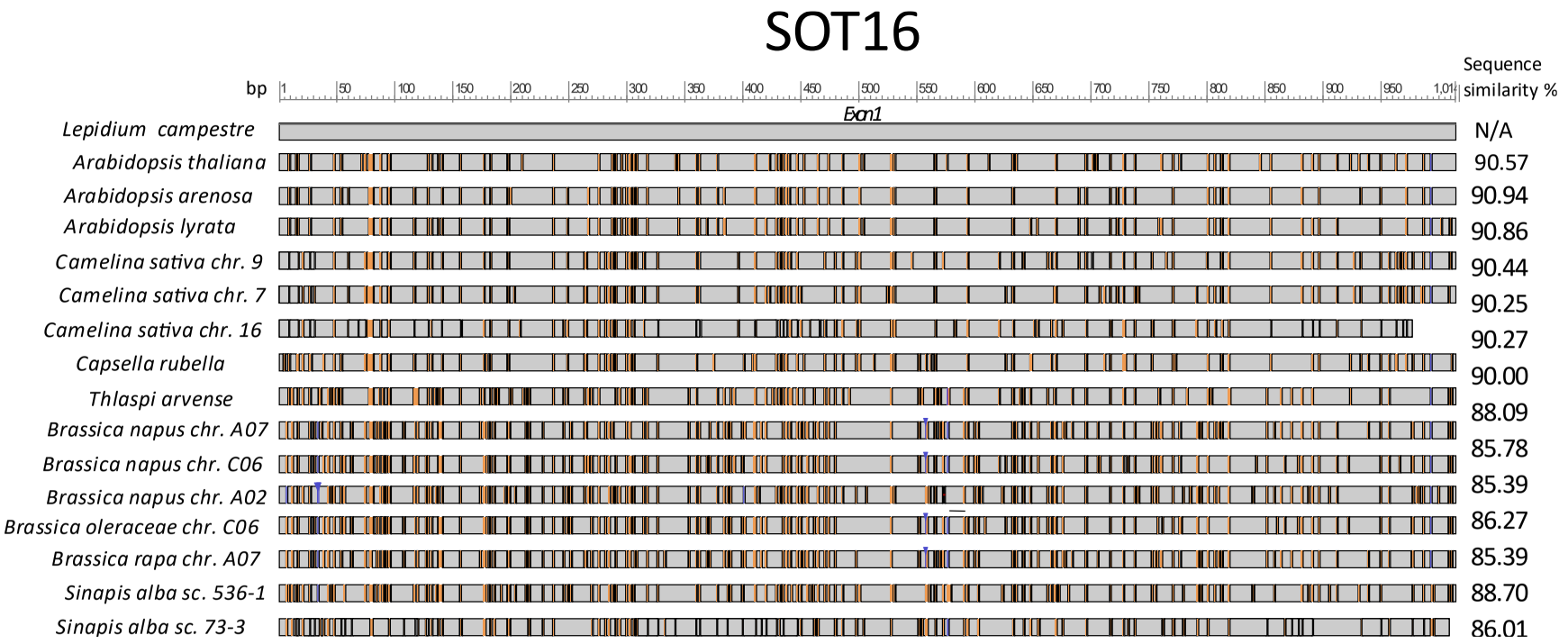


Figure 11. Visualization of *SOT16* gene conservation and sequence similarity across selected Brassicaceae species compared to *L. campestre*. Orange indicates a polymorphism, and blue indicates an insertion.

3.2 Evaluating polymorphisms found in *L. campestre*

Two sequencing-based methods, Sanger sequencing and targeted NGS, were applied to identify polymorphic loci in genes known to be involved in the regulation of the target traits. The aim being to identify desirable associations between loci and trait for high oil content, high OA content, low EA content, low GL content, low Sb and low Gla content.

3.2.1 Polymorphisms identified through Sanger sequencing

For this investigation, primers were designed to specifically target the coding regions of the four genes selected for an comparative genomics analysis. After optimization of the primers 25 *L. campestre* individuals were used for PCR amplification and consequetive Sanger sequencing. In total 1375 bp for *FAD2*, 1521 bp for *FAE1*, 1409 bp for *SOT16* and 2471 bp for *PEN2* were sequenced and after aligning the sequences from the 25 individuals six polymorphisms were found (Table 4). Three were identified in *FAD2*, two in *PEN2* and one in *SOT16*, while no polymorphisms were recorded for *FAE1*.1. Among these polymorphisms, one in a missense mutation (*FAD2*_16852827-C+2/-), leading to a codon deletion causing the loss of a serine amino acid.

Table 4. The six unique polymorphisms found across the four studied genes using Sanger sequencing. POS refers to position in its linkage group, and POS CDS to its bp position counted from ATG. A '+' refers to a polymorphism found downstream of CDS. A '*' in alternative allele refers to an ambiguous base, Y = C/T, M = A/C. Abbreviations: Alt. – Alternative, OA - Oleic acid, GLs - Glucosinolates, sb - sinalbin. Ref. – reference.

Polymorphism	Gene	Trait	Found in accessions	Phenotype of accessions	Ref. allele	Alt. allele	POS /POS CDS
*FAD2*_ 16852827-C+2/-	*FAD2*	Oil composition	88	High oil	CTT	-	16852827 / 31
*FAD2*_16853420-C/T	*FAD2*	Oil composition	88	High oil	C	Y*	16853420 / 624
*FAD2*_16853666-C/A	*FAD2*	Oil composition	88	High oil	C	M*	16853666 / 870
*PEN2*_1358072-C/G	*PEN2*	Glucosinolate content	92	Low OA, low GLs	C	G	1358072 / 864
*PEN2*_1359828-A/G	*PEN2*	Glucosinolate content	92, 94	Low OA, low GLs (92), low sb (94)	A	G	1359828 / +122
*SOT16*_18240573-G/C	*SOT16*	Glucosinolate content	94	Low OA, low GLs	G	C	18240573 / 9

3.2.2 Polymorphisms identified using targeted NGS

For the purpose of identifying additional polymorphisms in genes known to regulate the target traits, 15 genes were selected (Table 3) for targeted NGS sequencing. The gene sequences were retrieved from the *L. campestre* whole genome sequence data by using *A. thaliana* gene sequences as baits. In addition to the whole gene sequences, 200-300 bp up- and downstream of the genes were retrieved and used by CD genomics for designing a targeted NGS assay. Thirty-nine *L. campestre* and one *L. heterophyllum* individual was selected as a diversity panel based on previous data on seed oil content, oil composition and glucosinolate content and quality. High molecular weight genomic DNA was isolated from the diversity panel and sent for sequencing at CD genomics.

Table 5. The polymorphisms found across 15 studied genes (excluding accession LEP2020-313 and the polymorphism only recorded for this accession), shown as the number of polymorphisms in each gene as total, up-and downstream of the gene, introns and exons.

Gene	Trait	Total	Upstream of gene	Intron	Exon	Downstream of gene
KCS8	Oil quality	28	9	0	5	14
FAE1_1	Oil quality	0	0	0	0	0
FAE1_2	Oil quality	1	0	0	0	1
FAE1_3	Oil quality	16	8	0	5	3
FAD2	Oil quality	2	0	0	1	1
LEC1	Oil content	2	0	0	0	2
LEC2	Oil content	20	0	19	1	0
TAG1	Oil content	41	17	12	3	9
WRI1	Oil content	11	3	6	1	1
BUS1	Glucosinolate content	8	0	0	2	6
SOT16	Glucosinolate content	1	0	0	1	0
MYB28	Glucosinolate content	50	33	2	4	11
GTR1	Glucosinolate transport	16	10	0	4	2
GTR2	Glucosinolate transport	1	1	0	0	0
PEN2	Glucosinolate degradation	13	1	2	0	10
Total Polymorphisms:		210	82	41	27	60

Through alignments of the sequence data obtained from the targeted NGS sequencing a total of 671 polymorphic loci were identified. Accession Lep2020-313, which derives from a cross between *L. campestre x L. heterophyllum* had an alternative allele for 461 of these sites. Since this single accession captured more than half of all the identified polymorphic sites, these polymorphisms and this accession were excluded from further analysis as they were deemed to reflect the difference between *L. campestre* and *L. heterophyllum* rather than relevant polymorphic sites for *L. campestre*. After excluding LEP2020-313 and the 461 *L. heterophyllum*-related polymorphisms, 210 polymorphic sites derived from aligning 39 individuals were used for further analysis (Table 5). Of these, only 32

are present in one single accession, 72 are present in 2-3 accesions, 58 are present in 4-10 accessions, 13 occur in 11-20 accessions and 42 occur in 21-31 accessions.

The 27 polymorphisms found within exon regions were predicted for their effect on the final protein using the PolyPhen-2 software. The results showed that 19 of these polymorphisms were predicted to cause silent mutations, based on exon/intron structure predictions. Missense mutations were found in seven loci in oil related genes, and in six glucosinolate related genes (Table 6). Multiple of the mutations were predicted to have a detrimental effect on the protein function. The polymorphism *FAD2*_16852827-C+2/- was predicted to cause a codon deletion coding for a serine amino acid, which do not reside inside any protein domain (Table 6). Gustafsson *et al.* also found this deletion in their 2018 study of *L. campestre* genes.

Several of the missense mutations predicted to be detrimental to protein function, also showed significant correlations with oil or glucosinolate traits. *KCS8*_5299450-A/G is correlated with decreased oil content and EA, and increased GLs and OA (Table 6, Table 7). *MYB28*_13310330-A/C, *MYB28*_13311406-T/G and *MYB28*_13310279-T/C correlates with decreased oil content (Table 6, Table 7).

Interestingly – most missense mutations are predicted to have detrimental effects. However, majority of these mutations are not significantly linked to any phenotype.

Table 6. List of missense mutations and codon deletions found in oil- content and quality related genes of *L. campestre*. POS LG refers to its bp position in the linkage group, POS CDS to its bp position from ATG in the CDS. Codon refers to which codon is then affected counted from ATG as codon 1. Exon 7 * (*TAG1*) indicated that the exon is predicted with low support. Polyphen scores evaluate the probability of a substitution being damaging for protein. Values between 0-0.15 are predicted to be benign, 0.15-0.85 possibly damaging and 0.85-1 probably damaging.

Polymorphism	Gene	Trait	POS LG /POS CDS /Codon	Predicted Domain/loci functi (Interpro)	Exon	Ref. codon	Alt. codon	Codon property change	Statistically significantl linked?	Detrimental prediction (PolyPh 2)	Accessions with al allele
FAD2_ 16852827-C+2/-	FAD2	Oil quality - OA	16852827 /31 /11	PHOBIUS NON CYTOPLASMIC DOMAIN Predicted region of a membrane-bound, outside the membrane, in the extracell. region	Exon 1	Ser	-	Codon deletion	No	Probably damaging (score = 1.0)	3, 309, 310
FAE1-3_5986351-T/A	FAE1_3	Oil quality - VLCFA	5986351 /675 /225	N/A prediction could not be run	Exon 1	Asp	Glu	Larger	+ Gl (*) - Oil (*) + OA (*) - EA (*)	Probably damaging (score = 1.0)	309, 310
KCS8_5299450-A/G	KCS8	Oil quality - VLCFA	5299450 /143 /48	TMHMM - Region of a membrane-bound protein predicted to be embedded in the membrane.	Exon 1	Ile	Thr	Smaller, lost aliphatic, increased polarity,	+ Gl (*) - Oil (*) + OA (*) - EA (*)	Probably damaging (Score = 0.998)	3, 309, 310, 313
LEC2_3905387-C/T	LEC2	Oil content	3905387 781/261	MOBIDB_LITE mobidb-lite - consensus disorder prediction	Exon 5	Pro	Ser	Smaller, increased polarity	No	Benign (Score = 0.001)	3, 21, 23, 37, 47, 69, 70, 78, 80, 120, 126, 140, 180, 192, 194, 197, 198, 226, 234, 236, 262, 275, 298, 301, 302, 307, 309, 310, 313
TAG1_602932-G/T	TAG1	Oil content	602932 976/326	No	Exon 7 *	Val	Phe	Larger, lost aliphatic, gained aromatic	N/A	Uknown, score not possible	3
TAG1_602927-C/G	TAG1	Oil content	602927 971/324	No	Exon 7 *	Pro	Arg	Larger, increased polarity, gained positive	N/A	Uknown, score not possible	3
WRI1_16475133-T/C	WRI1	Oil content	16475133 886/296	COILS - Coil	Exon 6	Tyr	His	Gained positive	N/A	Possibly damaging (Score = 0.915)	3
GTR1_13757575-T/G	GTR1	Glucosinolate content	13757575 / 897/299	TMHMM TMhelix - Region of a membrane-bound protein predicted to be embedded in the membrane.	Exon 3	Asn	Lys	Larger, gained hydrophobic, gained positive	No	Benign (Score = 0.001)	3, 21, 23, 37, 47, 69, 70, 78, 80, 120, 126, 140, 180, 192, 194, 197, 198, 226, 234, 236, 262, 275, 298, 301, 302, 307, 309, 310, 313
GTR1_13758586-T/C	GTR1	Glucosinolate content	13758586 /202	TMHMM TMhelix - Region of a membrane-bound protein predicted to be embedded in the membrane.	Exon 4	Asn	Asp	Gained negative	No	Probably damaging (Score = 0.994)	89, 92
GTR1-13758088-T/C	GTR1	Glucosinolate content	13758088 /218	No	Exon 4	Lys	Glu	Lost hydrophobic, lost positive, gained negative	No	Possibly damaging (Score = 0.955)	89, 92
MYB28_13311406-T/G	MYB28	Glucosinolate content	13311406/ 854/285	No	Exon 3	Leu	His	Lost aliphatic, gained aromatic, gained polar, gained positive	- Oil (**)	Probably damaging (Score = 0.999)	69, 78, 298, 301, 302
MYB28_13310330-A/C	MYB28	Glucosinolate content	13310330/ 778/260	No	Exon 3	Asn	His	Larger, gained aromatic, gained hydrophobic, gained positive	- Oil (**)	Probably damaging (Score = 0.993)	69, 78, 298, 301, 302
MYB28_13310279-T/C	MYB28	Glucosinolate content	13310279/ 727/243	No	Exon 3	Phe	Leu	Lost aromatic, gained aliphatic	- Oil (**)	Benign (Score = 0.001)	309, 310

3.2.3 Significantly associated polymorphisms to target traits found using targeted NGS sequencing

Thirty nine accessions of *L. campestre* was sequenced in 15 genes associated with oil and GL content and composition to identify polymorphisms with significant association with these traits. In total 4778 bp of *WRI1*, 4338 bp of *TAG1*, 1233 bp of *LEC1*, 3546 bp of *LEC2*, 1691 bp of *FAD2*, 2140 bp of *FAE1_1*, 2058 bp of *FAE1_2*, 2005 bp of *FAE1_3*, 2133 bp of *KCS8*, 2196 bp of *MYB28*, 2319 bp of *BUS1*, 1924 bp of *SOT16*, 2803 bp of *PEN2*, 2662 bp of *GTR1* and 3174 bp of *GTR2* were sequenced.

In total 113 polymorphisms were found which were significantly associated with either oil, OA, EA, glucosinolate, Sb or Gl content, using targeted NGS sequencing (Table 7).

Markers for oil content in oil-content-related genes include the markers found in *WRI1*, *TAG1*, *LEC1* and *LEC2* (Table 7). Interestingly, most markers found within glucosinolate related genes was also shown to be significantly correlated with oil content rather than glucosinolates (Table 7). Several markers significantly associated with OA content were found in *WRI1* and and in *KCS8*, and a single in *LEC2* (*LEC2*_3905762-G/A) (Table 7). Moreover, there were additionally two markers identified in glucosinolate related genes –*BUS1* (*BUS1*_19870988-A/A+1) and *MYB28* (*MYB28*_13308312-G/A). Markers observed to be associated with increased OA also seem to be associated with either decreased oil content or increased GL. Eight markers were shown to be significantly associated with EA content markers and includes *TAG1*_604746-C/T, *LEC2*_3905849-C/T, *FAE1_2*_5978947-A/T, *GTR1*_13759152-G/C, *BUS1*_19870988-A/A+1, *MYB28*_13310279-T/C and all markers in *WRI1* and *KCS8*, respectively. The markers *TAG1*_604746-C/T and *BUS1*_19870988-A/A+1 are of most interest as they are associated with decreased EA but not decreased oil content (Table 7).

Polymorphisms of interest related to decreased glucosinolate content in glucosinolate related genes included *MYB28*_13310279-T/C and *GTR1*_13759152-G/C (Table 7). Polymorphisms with strong correlation to glucosinolate content were however also found in oil related genes such as in *WRI1* and *LEC2*. The *LEC2*_3905849-C/T marker was shown to associate favorably with glucosinolate decrease but unfavorably correlated with increased oil. For Sb and Gla content specifically, the markers are *MYB28*_13308312-G/A, *MYB28*_13311500-A/G, *BUS1*_19870988-A/A+1, *LEC2*_3905762-G/A, *TAG1*_604920-T/G and all markers in *WRI1* (Table 7).

There are a number of groups of polymorphic loci, often within close proximity within a linkage group, which are shared by the same accessions (Table 7), akin to haplotypes. The groups of polymorphisms can span multiple genes, as can be seen

in *FAD2* and *KCS8* (which are both located on LGX). This may be due to linkage disequilibrium causing higher than expected co-inheritance between these loci, or due to accessions being closely related to each other within homogenous populations.

The distribution of markers show that there is no single accession with all the benificial alleles for either oil content nor glucosinolate content (Table 8). The highest number of stacked markers for oil content is found in accession LEP2020 no. 70, 80, 88, 120, 126, 138, 180, 192, 197, 198, 222, 226, 234 and 236 all with ten out of 13 markers.

For OA content LEP2020-310 has the highest number of stacked markers, with three of total four found markers. For EA content LEP2020 no. 194 and 307 have all seven markers. For GL LEP2020 no. 89 and 92 have six out of ten markers. For Sb eight markers were found and seven for Gla (Table 8). Half of all accessions already possess all of the Gla markers, and seven of the Sb markers. Accessions with seven of both Sb and Gla markers are LEP2020 no. 21, 23, 37, 40, 70, 80, 88, 89, 92, 120, 138, 180, 192, 197, 222, 226 and 234 (Table 8).

Table 7. A list of all 113 polymorphisms observed in at least two accessions, with significant associations to traits of interest (OC, OA, EA, GL, Sb and Gla) in 39 studied accessions.

Markers are named according to [Gene], [Genome Position (POS)], and [Reference allele]/ [Alternative allele]. In cases of insertion/deletions only the first bp is written, with remaining shown as +X.

Colours in polymorphism name column indicate groupings where the same accessions share multiple alt. alleles. Trait refers to function associated with the gene the marker is in proximity of. POS refers to genomic bp position of polymorphism on Linkage group. P indicates significance interval, where $p \leq 0.05$ is considered significant. Positive association (green) indicate higher content with alt. allele, while negative (red) indicate higher content with Ref allele. Abbreviations: bp – base pair, GL – glucosinolate content, Sb – Sinalbin content, Gla – Glycoallysin content, OC – oil content, OA – oleic acid content, EA – erucic acid content, LG – Linkage group. VLCFA – Very long chain fatty acid content (oil composition).

Name of polymorphism	LG	Trait	REF	ALT	Gene Region	GL	Sb	Gla	OC	OA	EA
KCS8_5299934-C/T	LG3	VLCFA	C	T	Downstream	*			*	*	*
KCS8_5299919-T/C	LG3	VLCFA	T	C	Downstream	*			*	*	*
KCS8_5299917-A/G+1	LG3	VLCFA	A	GA	Downstream	*			*	*	*
KCS8_5299911-T/C	LG3	VLCFA	T	C	Downstream	*			*	*	*
KCS8_5299909-C/T	LG3	VLCFA	C	T	Downstream	*			*	*	*
KCS8_5299878-C/A	LG3	VLCFA	C	A	Downstream	*			*	*	*
KCS8_5299871-A/C+3	LG3	VLCFA	A	CTCA	Downstream	*			*	*	*
KCS8_5299859-A+21/T	LG3	VLCFA	ACAAAATAGTAAAACACATTT	-	Downstream	*			*	*	*
KCS8_5299829-T/G	LG3	VLCFA	T	G	Downstream	*			*	*	*
KCS8_5299815-G/A+1	LG3	VLCFA	G	AG	Downstream	*			*	*	*
KCS8_5299650-A/T	LG3	VLCFA	A	T	Downstream	*			*	*	*
KCS8_5299644-A/G	LG3	VLCFA	A	G	Downstream	*			*	*	*
KCS8_5299450-A/G	LG3	VLCFA	A	G	Exon	*			*	*	*
KCS8_5298555-T/C	LG3	VLCFA	T	C	Exon	*			*	*	*
KCS8_5298387-T/C	LG3	VLCFA	T	C	Exon	*			*	*	*
KCS8_5298357-T/A	LG3	VLCFA	T	A	Exon	*			*	*	*
KCS8_5298255-G/A	LG3	VLCFA	G	A	Exon	*			*	*	*
KCS8_5298096-A/G	LG3	VLCFA	A	G	Upstream	*			*	*	*
KCS8_5298093-A/G	LG3	VLCFA	A	G	Upstream	*			*	*	*
KCS8_5298062-A/T	LG3	VLCFA	A	T	Upstream	*			*	*	*
KCS8_5298022-C/A	LG3	VLCFA	C	A	Upstream	*			*	*	*
KCS8_5298009-A/-	LG3	VLCFA	A	-	Upstream	*			*	*	*
KCS8_5297866-G/A	LG3	VLCFA	G	A	Upstream	*			*	*	*
KCS8_5297801-T/C	LG3	VLCFA	T	C	Upstream	*			*	*	*
FAD2_16852827-C+2/-	LG3	OA	CTT	-	Exon	*			*	*	*
LEC1_1883961-C/T	LG1	OC	C	T	Intron				**		
LEC2_3903647-A/C	LG1	OC	A	C	Intron				*		
LEC2_3903968-A/G	LG1	OC	A	G	Intron				*		
LEC2_3904285-T/C	LG1	OC	T	C	Intron				*		
LEC2_3904362-C/T	LG1	OC	C	T	Intron				*		
LEC2_3904375-C/T	LG1	OC	C	T	Intron				*		
LEC2_3904380-C/T	LG1	OC	C	T	Intron				*		
LEC2_3904409-T/C	LG1	OC	T	C	Intron				*		
LEC2_3904572-A+15/T	LG1	OC	AAATATTAATTTATAT	T	Intron				*		
LEC2_3904738-T+2/-	LG1	OC	TTC	-	Intron				*		
LEC2_3905307-T/A	LG1	OC	T	A	Intron				*		
LEC2_3905423-T/G	LG1	OC	T	G	Intron				*		
LEC2_3905680-G/T	LG1	OC	G	T	Intron				*		
LEC2_3905762-G/A	LG1	OC	G	A	Intron	**	**	**		**	
LEC2_3905801-T+2/-	LG1	OC	TTA	-	Intron				*		
LEC2_3905804-C/T	LG1	OC	C	T	Intron				*		
LEC2_3905844-A+2/-	LG1	OC	ATA	-	Intron				*		
LEC2_3905849-C/T	LG1	OC	C	T	Intron	*			**		**
LEC2_3906180-A/G	LG1	OC	A	G	Intron				*		
TAG1_604920-T/G	LG3	OC	T	G	Downstream		*	*	**		
TAG1_604746-C/T	LG3	OC	C	T	Downstream						**
TAG1_603528-T/C	LG3	OC	T	C	Intron				**		
TAG1_602057-A/G	LG3	OC	A	G	Intron				**		
TAG1_600820-A/G	LG3	OC	A	G	Upstream						
TAG1_600844-A/G	LG3	OC	A	G	Upstream				**		
TAG1_600842-A/G	LG3	OC	A	G	Upstream				**		
TAG1_600819-A/G	LG3	OC	A	G	Upstream				**		
TAG1_600784-A/G	LG3	OC	A	G	Upstream				**		
TAG1_600759-T/C	LG3	OC	T	C	Upstream				**		
TAG1_600729-T/C	LG3	OC	T	C	Upstream				**		
FAE1_2_5978947-A/T	LG5	OC	A	T	Downstream				*		*
FAE1_3_5985347-A/A+3	LG5	OC	A	AAAT	Upstream						
FAE1_3_5987325-C/T	LG5	OC	C	T	Downstream				*		
WRI1_16471364-C/C+1,C+2	LG5	OC	C	CA,CAA	Upstream	*	**	**	***	*	*
WRI1_16472916-G/G+1	LG5	OC	G	GA	Intron	**	**	***	***	**	**
WRI1_16473045-C/T	LG5	OC	C	T	Intron	**	**	***	***	**	**
WRI1_16473352-C/C+7	LG5	OC	C	CTTTATTT	Intron	*			*		*
WRI1_16475686-T/C	LG5	OC	T	C	Downstream				**		*

	Alt > Ref	Alt < Ref
p = >0.075		
p = 0.05-0.075		
p = 0.01-0.05	*	*
p = 0.001-0.01	**	**
p = 0.0001-0.001	***	***

Name of polymorphism	LG	Trait	REF	ALT	Gene Region	GL	Sb	Gla	OC	OA	EA
BUS1_19870869-C+11/C	LG1	GL	CAACTATAATAA	C	Downstream				**		
BUS1_19870988-A/A+1	LG1	GL	A	AC	Downstream		*			*	**
PEN2_1360052-T/A	LG4	GL	T	A	Downstream						
PEN2_1359828-G/A	LG4	GL	G	A	Downstream				*		
GTR1_13759152-G/C	LG5	GL	G	C	Downstream	*			*		*
MYB28_13311500-A/G	LG8	GL	A	G	Downstream		*	*	*		
MYB28_13311490-G+21/C	LG8	GL	GACTTCAAACCTTGACTCCTC	C	Downstream				**		
MYB28_13311440-G/C	LG8	GL	G	C	Downstream				**		
MYB28_13311409-T/C	LG8	GL	T	C	Downstream				**		
MYB28_13311406-T/G	LG8	GL	T	G	Downstream				**		
MYB28_13311385-G/A	LG8	GL	G	A	Downstream				**		
MYB28_13311372-A/T+2	LG8	GL	A	TTA	Downstream				**		
MYB28_13311370-A/T+6	LG8	GL	A	TATAAAA	Downstream				**		
MYB28_13311352-C/A+1	LG8	GL	C	AC	Downstream				**		
MYB28_13310847-T/C	LG8	GL	T	C	Downstream				**		
MYB28_13310406-T/A	LG8	GL	T	A	Exon				**		
MYB28_13310344-G/A	LG8	GL	G	A	Exon				**		
MYB28_13310330-A/C	LG8	GL	A	C	Exon				**		
MYB28_13310279-T/C	LG8	GL	T	C	Exon	*			*		*
MYB28_13309725-C/T	LG8	GL	C	T	Intron				**		
MYB28_13309491-C/T	LG8	GL	C	T	Intron				**		
MYB28_13309251-C+12/A	LG8	GL	CAATATATATATA	A	Upstream				**		
MYB28_13308864-A/T	LG8	GL	A	T	Upstream				**		
MYB28_13308767-C/T	LG8	GL	C	T	Upstream				**		
MYB28_13308750-A/G	LG8	GL	A	G	Upstream				**		
MYB28_13308743-C/T	LG8	GL	C	T	Upstream				**		
MYB28_13308739-A/G	LG8	GL	A	G	Upstream				**		
MYB28_13308732-G/A	LG8	GL	G	A	Upstream				**		
MYB28_13308730-G/A	LG8	GL	G	A	Upstream				**		
MYB28_13308714-G/A	LG8	GL	G	A	Upstream				**		
MYB28_13308704-A/G	LG8	GL	A	G	Upstream				**		
MYB28_13308684-T/C	LG8	GL	T	C	Upstream				**		
MYB28_13308680-A/C	LG8	GL	A	C	Upstream				**		
MYB28_13308673-G/A	LG8	GL	G	A	Upstream				**		
MYB28_13308653-T/C	LG8	GL	T	C	Upstream				**		
MYB28_13308648-G/A	LG8	GL	G	A	Upstream				**		
MYB28_13308644-A/G	LG8	GL	A	G	Upstream				**		
MYB28_13308636-T/A	LG8	GL	T	A	Upstream				**		
MYB28_13308606-T/A	LG8	GL	T	A	Upstream				**		
MYB28_13308604-G/A	LG8	GL	G	A	Upstream				**		
MYB28_13308590-G/A	LG8	GL	G	A	Upstream				**		
MYB28_13308502-A/C	LG8	GL	A	C	Upstream				**		
MYB28_13308494-C/A	LG8	GL	C	A	Upstream				**		
MYB28_13308485-T/C	LG8	GL	T	C	Upstream				**		
MYB28_13308478-C/T	LG8	GL	C	T	Upstream				**		
MYB28_13308379-A/G	LG8	GL	A	G	Upstream				**		
MYB28_13308376-C/T	LG8	GL	C	T	Upstream				**		
MYB28_13308351-C/T	LG8	GL	C	T	Upstream				**		
MYB28_13308346-T/A	LG8	GL	T	A	Upstream				**		
MYB28_13308312-G/A	LG8	GL	G	A	Upstream		**	**		**	

Table 8. Distribution of marker groups with significant correlation to oil (A) or glucosinolate content (B), for each accession. No. In Group indicate the number of polymorphic loci in a group (colour coded as in Table 7, column "name of polymorphism") which are shared by the same accessions. Trait association indicate which trait the marker is significant for, while R indicates that reference allele is preferred and A that alternative allele is preferred.

Abbreviations: EA – Erucic acid, GL – Glucosinolates, Gla – Glucoallysin, OA – oleic acid, OC – Oil content, Sb – Sinalbin.

A	Significant oil-related markers																B	Significant glucosinolate-related markers																	
no. In Group	2	1	1	16	8	44	1	1	25	1	1	2	1	1	1	1		2	1	1	25	1	1	1	2	1	1	1	1	1	1			
Trait association /Beneficial Ref. or Alt. allele	OC/R	OC/R	OC/R	OC/A	OC/R	OC/A	OC/R	OC/A	OC/R OA/A EA/R	OC/A EA/R	OC/A EA/R	OC/R EA/A	OC/R EA/A	OA/A EA/A	OA/A OA/A	EA/A		GL/R	GL/R Sb/R Gla/R	GL/A	GL/R	GL/R	GL/R	GL/R Sb/R Gla/R	GL/R Sb/R Gla/R	Sb/R Gla/R	Sb/R Gla/R	Sb/R Gla/R	Sb/R	Sb/R				
Marker name (right) Acc. No (down)	WRI1_16472916-G/GA *	WRI1_16471364-C/CA,CAA	TAG1_604920-T/G	LEC2_3903647-A/C	TAG1_603528-T/C *	LEC1_1883961-C/T *	BUS1_19870869-AACTATAATAAA/-	FAE1_3_5987325-C/T	FAD2_16852827-CTT/-*	LEC2_3905849-C/T	FAE1_2_5978947-A/T	WRI1_16473352-C/CTTTATTT *	WRI1_16475686-T/C	BUS1_19870988-A/AC	LEC2_3905762-G/A	TAG1_604746-C/T		GTR1_13759152-G/C*	LEC2_3905762-G/A	LEC2_3905849-C/T	KCS8_5299934-C/T*	PEN2_1360052-T/A	PEN2_1359828-G/A	WRI1_16471364-C/CA,CAA	WRI1_16472916-G/GA*	MYB28_13311500-A/G	TAG1_604920-T/G	MYB28_13308312-G/A	BUS1_19870988-A/AC	WRI1_16475686-T/C				
LEP2020-3		X	X	X	X								X									X		X		X	X	X	X	X			
LEP2020-21		X	X		X		X	X	X	X	X		X							X	X	X		X	X	X	X	X	X	X			
LEP2020-23		X	X		X		X	X	X	X	X		X							X	X	X		X	X	X	X	X	X	X			
LEP2020-37		X	X		X		X	X	X	X	X		X							X	X	X		X	X	X	X	X	X	X			
LEP2020-40		X	X		X		X	X	X	X	X		X							X	X	X		X	X	X	X	X	X	X			
LEP2020-47			X		X		X	X	X	X	X			X		X				X	X						X						
LEP2020-69				X		X			X												X							X	X					
LEP2020-70	X	X	X		X		X	X	X	X	X		X							X	X	X		X	X	X	X	X	X	X			
LEP2020-72	X	X	X	X	X		X		X				X								X	X	X	X	X	X	X	X	X	X			
LEP2020-78				X		X			X												X							X	X					
LEP2020-80	X	X	X		X		X	X	X	X	X		X							X	X	X		X	X	X	X	X	X	X			
LEP2020-81	X	X	X	X	X		X		X				X								X	X	X	X	X	X	X	X	X	X			
LEP2020-88	X	X	X		X		X	X	X	X	X		X							X	X	X		X	X	X	X	X	X	X			
LEP2020-89	X	X	X		X		X		X	X			X							X	X	X	X	X	X	X	X	X	X	X			
LEP2020-92	X	X	X		X		X		X	X			X							X	X	X	X	X	X	X	X	X	X	X			
LEP2020-94	X	X	X	X	X		X		X				X								X	X	X	X	X	X	X	X	X	X			
LEP2020-120	X	X	X		X		X	X	X	X	X		X							X	X	X		X	X	X	X	X	X	X			
LEP2020-126	X	X	X		X		X	X	X	X	X		X	X						X	X	X		X	X	X	X	X		X			
LEP2020-138	X	X	X		X		X	X	X	X	X		X							X	X	X		X	X	X	X	X	X	X			
LEP2020-140	X	X			X		X		X				X	X							X	X		X		X		X		X			
LEP2020-180	X	X	X		X		X	X	X	X	X		X							X	X	X		X	X	X	X	X	X	X			
LEP2020-192	X	X	X		X		X	X	X	X	X		X							X	X	X		X	X	X	X	X	X	X			
LEP2020-194			X		X		X		X					X		X					X						X						
LEP2020-197	X	X	X		X		X	X	X	X	X		X							X	X	X		X	X	X	X	X	X	X			
LEP2020-198	X	X	X		X		X	X	X	X	X		X	X						X	X	X		X	X	X	X	X		X			
LEP2020-222	X	X	X		X		X	X	X	X	X		X							X	X	X		X	X	X	X	X	X	X			
LEP2020-226	X	X	X		X		X	X	X	X	X		X							X	X	X		X	X	X	X	X	X	X			
LEP2020-234	X	X	X		X		X	X	X	X	X		X							X	X	X		X	X	X	X	X	X	X			
LEP2020-236	X	X	X		X		X	X	X	X	X		X							X	X	X		X	X	X	X	X	X	X			
LEP2020-257	X	X			X		X		X				X	X							X	X		X		X		X		X			
LEP2020-262							X		X	X				X						X	X		X										
LEP2020-275							X		X	X				X						X	X		X										
LEP2020-298				X		X			X												X							X	X					
LEP2020-301				X		X			X												X							X	X					
LEP2020-302				X		X			X												X							X	X					
LEP2020-303			X		X		X		X				X		X				X		X					X		X	X					
LEP2020-307			X		X		X		X						X	X			X		X					X			X					
LEP2020-309				X			X					X	X	X				X									X		X		X			
LEP2020-310				X			X					X		X		X		X									X		X					

3.2.4 Principle Component Analysis (PCA)

A PCA was performed on the set of 210 polymorphisms found in the targeted NGS-sequencing, to evaluate if the genetic variation could be explained by observed phenotype. Components 1 and 2 describe 24.2% and 29.5% of the variation, respectively, with three notable clusters of which one contains the

majority of data points. The PCA plots shows no clear clustering pattern of accessions of similar phenotype (Figure 12A and B). This indicates that the *L. campestre* accessions have little genetic variation, which is also supported by prior observations (Geleta *et al.* 2020). There are some accessions distinctly grouping together – however these do not share phenotypic traits (Figure 12B), instead some of the observed variation can be explained by geographical origin (Figure 12C).

PCA-analysis was also performed separately for all polymorphisms identified in or in proximity to genes known to regulate oil content, oil composition and glucosinolate content, respectively. The results from these analyses yielded results very similar to Figure 12 and are therefore not shown.

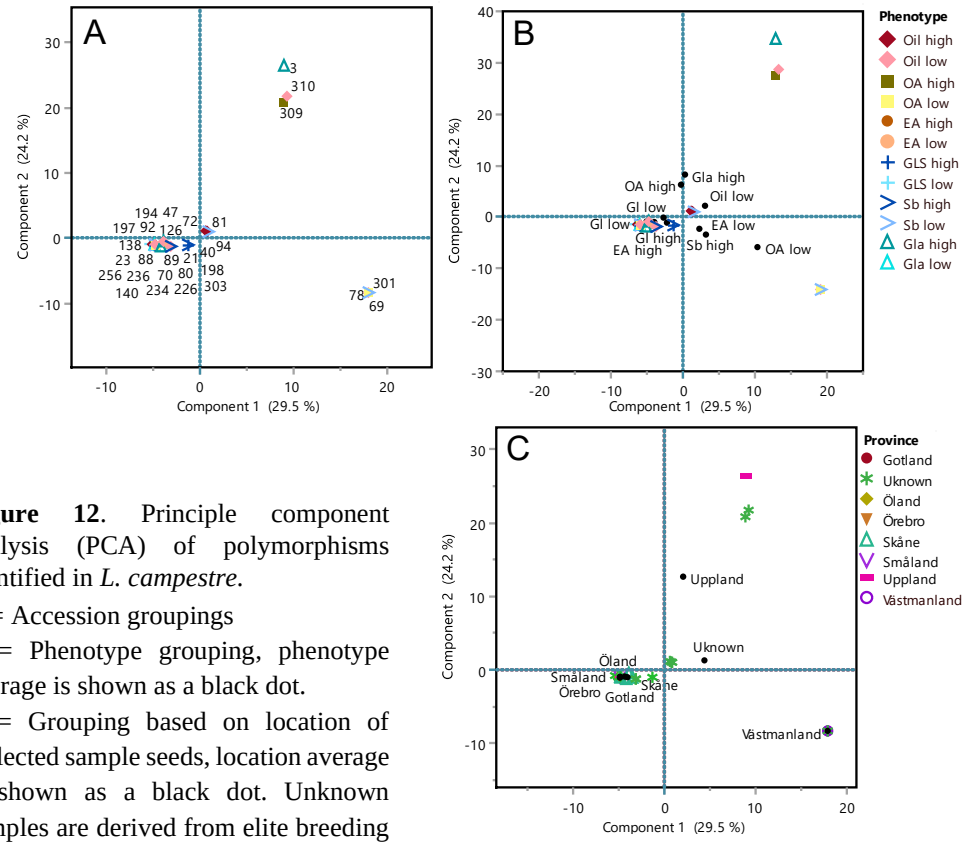


Figure 12. Principle component analysis (PCA) of polymorphisms identified in *L. campestre*.

A = Accession groupings

B = Phenotype grouping, phenotype average is shown as a black dot.

C = Grouping based on location of collected sample seeds, location average is shown as a black dot. Unknown samples are derived from elite breeding lines of mixed origin.

3.3 Method comparison – Targeted NGS and Sanger sequencing

Since two different methods for identifying polymorphisms were used in this study it gives an opportunity to compare the two methods. Sanger sequencing were performed at Eurofins Genomics for 25 individuals and four genes divided into 11 amplicons, summarizing to 7190 bp. NGS sequencing was performed by CD-genomics on pre-extracted DNA from 40 individuals and 15 genes, summarizing to roughly 43190 bp.

Evaluation of the number of polymorphisms found in Sanger versus targeted NGS sequencing

As evident by this study (Table 5) and supported by prior findings (Ching *et al.* 2002, Ghosh et al. 2002), a higher number of polymorphisms are found outside of the coding region rather than within exons.

There was a difference in the number of polymorphisms discovered using Sanger (6) and NGS (210) sequencing in this study. This is however obviously due to the lower number of genes, samples, and lower number of base pairs in genes that was sequenced using Sanger compared to targeted NGS.

The same polymorphisms were found in the same region, however observed in different accessions. In *FAD2*, three polymorphisms were found in LEP2020-88 that were not observed with targeted NGS. It was further observed that *FAD2*_16852827-CTT/- was found in LEP2020-89 with NGS, but not with Sanger. The first among two SNPs found in *PEN2* in Sanger sequencing (*PEN2*_1358072-C/G) did occur within the same accessions using Sanger and NGS. The second however (*PEN2*_1359828-A/G) had two accessions where the SNP was observed using Sanger and not with NGS, and nine accessions where the SNP was found using NGS and not with Sanger. *SOT16*_18240573-G/C was observed in and six accession in NGS, and none in Sanger.

Evaluation of cost and time management with Sanger versus NGS sequencing

In terms of cost, targeted NGS was unquestionably less costly than the Sanger sequencing when the actual NGS and theoretical Sanger cost of 15 genes was compared. A lower cost per sequenced gene and polymorphism was identified with NGS (Table 9). High shipping costs (284 USD) with CD-genomics offsets this monetary advantage when working with very few PCR-products. In these cases, the Sanger sequencing may be preferred. Sanger sequencing by Eurofins was carried out in tubes or plate format, with pre-paid barcode labels. The cheaper Sanger sequencing plate alternatives comes with a minimum number of samples in 96-well plate format.

From the calculations alone, it appeared as the most cost-effective sequencing alternative at Eurofins is sending unpurified PCR products in a 96-plate format to be purified and sequenced at their facilities (Table 9). However, this is misleading and does not consider the rate of failure for sending in unpurified versus pre-purified samples. It was observed in this study that when sending unpurified PCR products (in plates or tubes) for Sanger sequencing there was a higher likelihood of sequencing failure as well as an overall lower sequencing quality when successfully sequenced compared to when sending purified PCR product (purified in-house using a kit), – offsetting the cost-benefit. This may be due to the difficulties of properly estimating PCR-concentrations prior purification, or differences in the purification performed at SLU and Eurofins. Another cost-related issue not included in the calculation is the cost of in-house labour for the Sanger sequencing, which is substantial – estimated weeks-months of full-time work for a 15-gene study. Furthermore, primer-pair optimization, sequencing failure and other possible costly mishaps are not accounted for. Neither is potential licences for alignment software, PCR, Nanodrop, gel electrophoresis chamber and gel documentation and camera equipment and maintenance. DNA-extraction related costs necessary for both Sanger and NGS was also excluded from the calculation.

Table 9. Cost (in SEK) for sequencing and identification of polymorphisms in the 15 genes included in this study by the targeted Next-Generation Sequencing (NGS) (CD-genomics), and the theoretical cost of using Sanger sequencing (at Eurofins) for purified Polymer Chain Reaction (PCR) products in tubes, plate or unpurified products in plate format. Costs for Sanger sequencing were calculated by estimates based on data from 15 genes and 40 accessions in this study with average gene length of 2080 bp +-700 bp up and downstream and five primer pairs per gene, including cost of PCR and sequencing-related products and services. Polymorphism refers to a single locus with a polymorphism in at least one accession (Table 7). Exchange rate used was 1 USD = 10.40 SEK.

Cost per	Sanger (Tube - Purified)	Sanger (Purified plate)	Sanger (Unpurified plate)	Targeted NGS
Per sequenced Bp	5.69	3.85	3.69	1.57
Gene	19803.22	13395.25	12830.48	5466.70
Polymorphism	1414.52	956.80	916.46	390.48

In terms of time, compared to targeted NGS, Sanger sequencing, as already stated, require vastly more manual work for the individual researcher – as the PCR, purification and gel electrophoresis must be done manually for all samples. Another drawback to consider for Sanger is the quality of sequencing which will likely be poorer due to less coverage depth. In this study, sequencing was performed once in the forward and once in the reverse direction. A study of this scale (sequencing of 15 genes including up- and downstream sequences, for 40 individuals) would have been impossible to achieve within this project's time frame using only Sanger sequencing – mainly due to time constraints. A proper experimental setup is more crucial for targeted NGS approaches due to the high shipment costs and long

waiting time for results, necessitating larger batch shipments, which may be an issue for some studies. NGS also require vastly more prior genomic annotation, putting it out of reach for many novel studies.

3.4 Evaluation of traits in *Lepidium* species for future interspecific breeding with *Lepidium campestre*

3.4.1 Evaluation of selected polymorphisms in oil content related genes in *Lepidium* species

Of the 113 polymorphisms initially discovered in *L. campestre* (Table 7), 14 loci with significant association to oil content or composition in oil-related genes were chosen for further evaluation in 15 *Lepidium* accessions of 11 different species (Table 10). Leaves for sampling could not be obtained for *L. cardamines* where therefore excluded from analysis. Polymorphisms will be referred to in the order they appear in the amplicon sequence.

Table 10. Polymorphisms discovered *in L. campestre* (Table 7) chosen for further evaluation in *Lepidium* accessions. Phenotypes marked in red indicates negative correlation, while green indicates positive. POS indicates the position of locus in relationship to the coding sequence, with – being upstream of ATG and + being downstream of stop-codon.

SNP Name	Statistically linked to:	POS	Gene	Region
*FAD2*_16852827-CTT/-	Oil content*, OA content *, EA content *	32	*FAD2*	Exon
*WRI1*_16471364-C/CA,CAA	Oil content***, OA content *, EA content *	-257	*WRI1*	Upstream
*WRI1*_16472916-G/GA	Oil content***, OA content **, EA content **	1296	*WRI1*	Intron
*WRI1*_16473045-C/T	Oil content***, OA content **, EA content **	1425	*WRI1*	Intron
*WRI1*_16473352-C/CTTTATTT	Oil content*, EA content *	1732	*WRI1*	Intron
*WRI1*_16475686-T/C	Oil content**, EA content *	+168	*WRI1*	Downstream
*TAG1*_602057-A/G	Oil content**	736	*TAG1*	Intron
*TAG1*_604746-C/T	EA content **	+323	*TAG1*	Downstream
*TAG1*_604920-T/G	Oil content**	+497	*TAG1*	Downstream
*LEC2*_3905801-TTA/-	Oil content**	2903	*LEC2*	Intron
*LEC2*_3905804-C/T	Oil content*	2879	*LEC2*	Intron
*LEC2*_3905844-ATA/-	Oil content*	2852	*LEC2*	Intron
*LEC2*_3905849-C/T	Oil content*	2858	*LEC2*	Intron
*LEC2*_3906180-A/G	Oil content*	3234	*LEC2*	Intron

FAD2

As previously demonstrated in the comparative genomic study of species from the Brassicaceae (Figure 7) the *FAD2* gene conservation structure is high. No INDELs were observed between the species. *L. campestre* and *L. heterophyllum* even share an identical sequence. It is also evident that the *L. hirtum* subspecies have a high sequence similarity, with only 1-2 substitutions. Highly polymorphic regions that were consistent across all species except *L. heterophyllum* and *L. hirtum*, such as at 169 bp, 241 bp, 273 bp, 378 bp and 456 bp were observed (Figure 13). The targeted SNP (*FAD2*_16852827-C+2/-) was shown to be

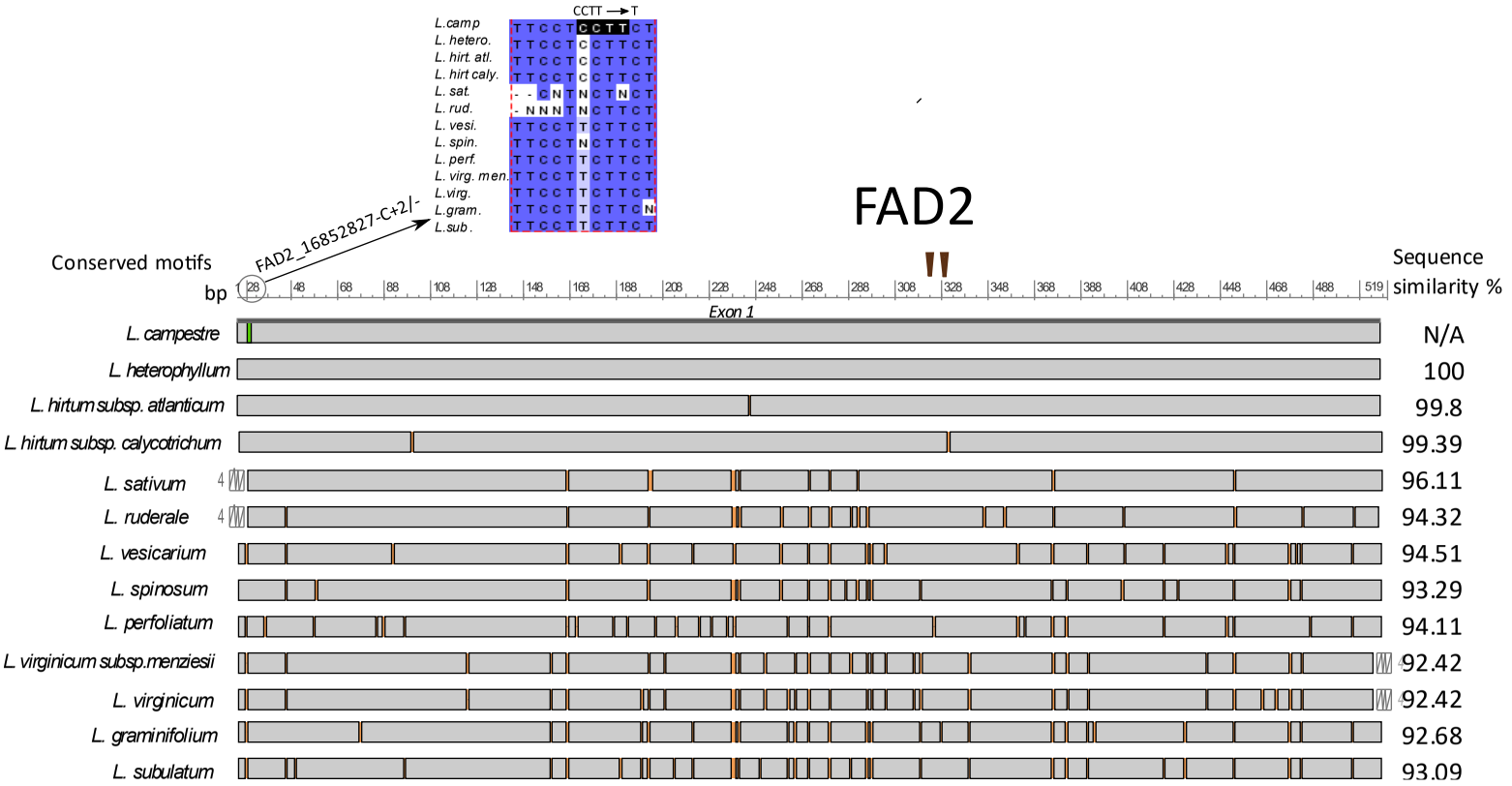


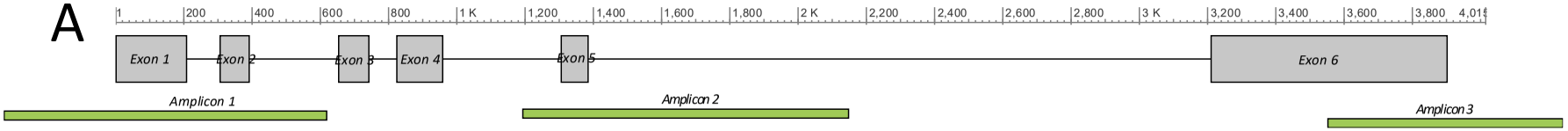
Figure 13. Visualization of *FAD2* gene alignment for 12 *Lepidium* species in positions 28-519 bp. A band in the *L. campestre* sequence indicates where an interesting polymorphism has been found in *L. campestre*, an orange band in the *Lepidium* accessions indicates substitutions compared to *L. campestre*. Region around the target polymorphism (*FAD2*_16852827-C+2/-) shown in detail, in blue. A yellow arrow indicates where a conserved motif is located.

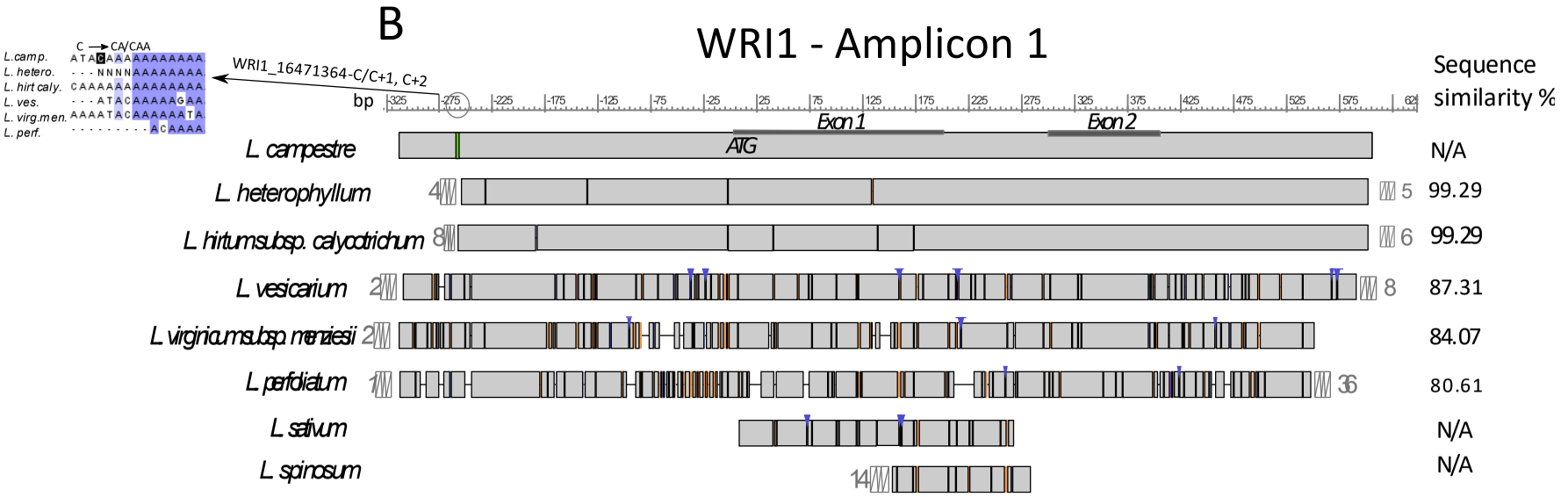
polymorphic across all species except *L. heterophyllum, L. hirtum* and *L. sativum* (Figure 13), with the CCTT substituted for T, which subsequently substituted the amino acid proline for serine. The allele in *L. campestre* result in a deletion of a serine amino acid.

WRI1

The results from sequencing the three *WRI1* gene regions holding the five target SNPs (Figure 14A) showed that there is a lower level of conservation in *WRI1* compared to that of *FAD2*. In this region of *WRI1*, there are multiple gaps and insertions present in all three studied amplicons (Figure 14B-D). INDELs and substitutions are especially prevalent directly adjacent to exon/intron junctions but are also present frequently within exons (Figure 14B-D). The 200 bp region flanking exon six downstream is interestingly relatively well conserved (Figure 13D). The position of the first targeted SNP of interest (*WRI1*_16471364-C/C+1,C+2) in amplicon 1 was not successfully sequenced in most species. However, in *L. hirtum subsp. calycotrichum, L. virginicum subsp. menziesii* and *L. vesicarium*, a substitution from C to T was recorded (Figure 14B). Amplicon 2 had overall lower sequencing success-rate, despite two attempts with different primers, possibly due to it being a highly polymorphic region. The first SNP (*WRI1*_16472916-G/G+1) in amplicon 2 is not present in *L. heterophyllum* due to its placement in a region where *L. heterophyllum* has a deletion (Figure 14C). The position in *L. vesicarium* and *L. spinosum* is conserved with *L. campestre* (Figure 14C).

The second SNP (*WRI1*_16473045-C/T) in amplicon 2 is conserved between *L. campestre* and *L. heterophyllum* with a substitution in *L. vesicarium* and *L. spinosum* (Figure 14C). The third polymorphism in amplicon 2 (*WRI1*_16473352-C/C+7) is causing ann insertion of 7 bp in a loci that is conserved across all species, in a region that is AT-rich (Figure 14C). The SNP in amplicon 3 (*WRI1*_16475686-T/C) is also in a conserved region (Figure 14D).


A
1
200
400
600
800
1 K
1,200
1,400
1,600
1,800
2 K
2,200
2,400
2,600
2,800
3 K
3,200
3,400
3,600
3,800
4,01
Exon 1
Exon 2
Exon 3
Exon 4
Exon 5
Exon 6
Amplicon 1
Amplicon 2
Amplicon 3


B
WRI1 - Amplicon 1
C → CA/CAA
L.camp. ATACAAAAAAAAAAA
L. hetero. - - - NNNNAAAAAAAA
L. hirt caly. CAAAAAAAAAAAAAA
L. ves. - - - ATACAAAAAGAA
L. virg.men. AAAATACAAAAAATA
L. perf. - - - - - - - - - ACAAAA
WRI1_16471364-C/C+1, C+2
bp
-325
-275
-225
-175
-125
-75
-25
25
75
125
175
225
275
325
375
425
475
525
575
625
Exon 1
Exon 2
ATG
L. campestre
L. heterophyllum
L. hirtumsubsp. calycotrichum
L. vesicarium
L. virginicumsubsp. menziesii
L. perfoliatum
L. sativum
L. spinosum
Sequence similarity %
N/A
99.29
99.29
87.31
84.07
80.61
N/A
N/A

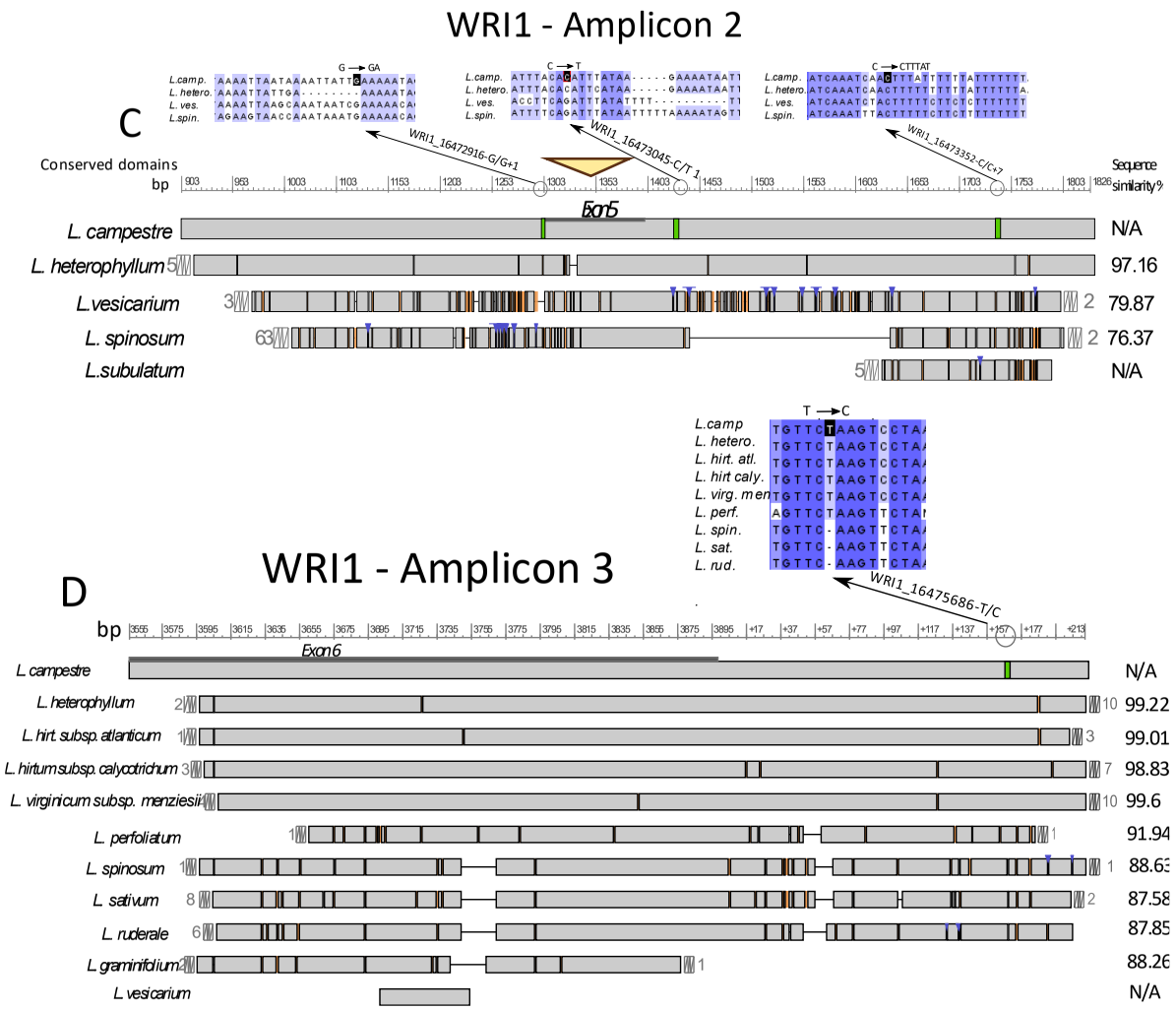


Figure 14. (A) - Visualization of the predicted *WRI1* structure based on *Arabidopsis* TAIR data, with the three sequenced amplicons shown in green. Gene alignment in the studied regions between -325-625 bp (Amplicon 1) (B), 903-1826 bp (Amplicon 2) (C) and 3555-+213 bp (Amplicon 3) (D), position relative to CDS from ATG. A band in the *L. campestre* sequence indicates an interesting polymorphism, an orange band in the *Lepidium* accessions indicates substitutions compared *to L. campestre*, and a blue band an insertion with an arrow for longer insertions. A yellow arrow indicates where a conserved domain is located. Sequence similarity is not included when amplicons are <50% coverage to reference. Regions around the target polymorphisms (*WRI1*_16471364-C/C+1,C+2, *WRI1*_16472916-G/G+1, *WRI1*_16473045-C/T, *WRI1*_16473352-C/C+7 and *WRI1*_16475686-T/C) shown in detail, in blue.

TAG1

Two regions in *WRI1* holding three target SNPs were evaluated in *Lepidium* accessions (Figure 15A). The results from sequencing the two amplicons showed that *TAG1* is relatively well conserved within the exons, with prevalence of INDELs present in the intronic regions (Figure 15). A conserved deletion can be found in the second intron in *L. virginicum, L. graminifolium* and *L. subulatum*, and three in the third intron in *L. graminifolium* and *L. subulatum* in amplicon 1 (Figure 14B). A highly polymorphic region was found in the first half of intron 3, in all species except *L. heterophyllum* and *L. hirtum* (Figure 15B). The locus of the first SNP of interest in amplicon 1 (*TAG1*_602057-A/G) is generally conserved in most of the studied species, except *L. campestre* and *L. subulatum* (Figure 15, B). Noteworthy, the substitution changes A to G, while G is the nucleotide found across almost all the other species (Figure 15B). The second SNP (*TAG1*_604746-C/T) of interest in amplicon 2 was shown to be very well conserved. The third SNP (*TAG1*_604920-T/G) is in a locus with the substitution from T to G in *L. campestre*, while all other studied species have a G (Figure 15B-C).

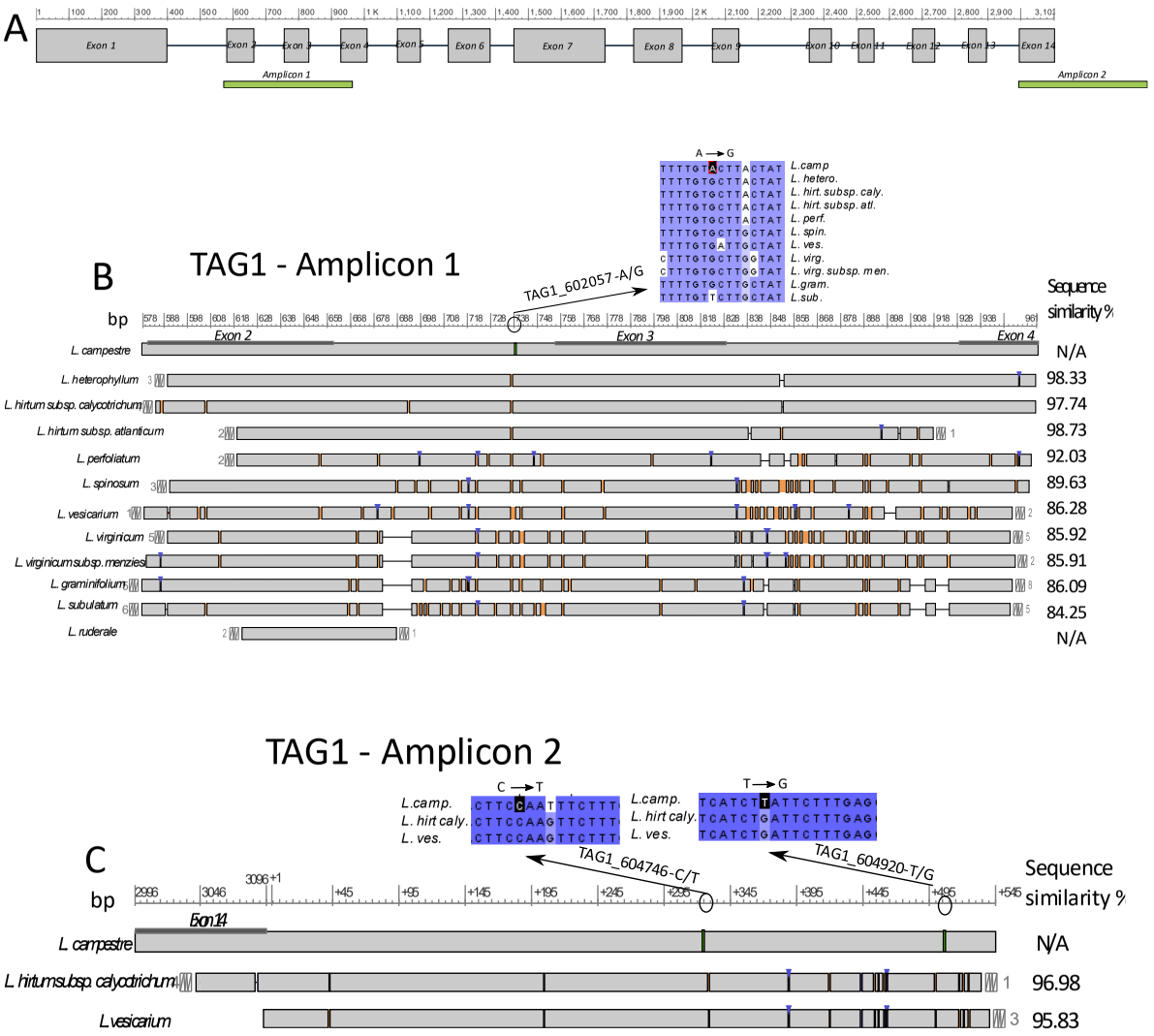


Figure 15. (A) - Visualization of the predicted *TAG1* structure based on *Arabidopsis* TAIR data, with the two sequenced amplicons shown in green. Gene alignment in the studied regions between 578-961 bp (Amplicon 1) (B) and 2996-+545 bp (Amplicon 2) (C), position relative to CDS from ATG. A band in *L. campestre* indicates an interesting SNP, an orange band in the *Lepidium* accessions indicates substitutions compared to *L. campestre*, a blue band an insertion with an arrow for longer insertions Sequence similarity is not included when amplicons are <50% coverage to reference. Regions around the target polymorphisms (*TAG1*_602057-A/G, *TAG1*_604746-C/T and *TAG1*_604920-T/G) shown in detail, in blue.

LEC2

A region in *LEC2* holding five target polymorphisms were evaluated in *Lepidium* accessions (Figure 16). The results from sequencing show that the level of conservation was high in *LEC2* for some species, including the closely related *L. hirtum,* and more distant *L. virginicum* (Figure 16). Interestingly, *L. heterophyllum* was not shown to have a high sequence similarity to *L. campestre*, with a unique deletion in intron 5 and a highly polymorphic region not shared with other *Lepidium* (Figure 16). Most of the intronic region in intron 5 is highly polymorphic and could not be properly aligned in *L. perfoliatum* and *L. ruderale*. The same region was not successfully sequenced in *L. virginicum, L. spinosum, L. subulatum* and *L. graminifolium*, clearly demonstrating that this intron has a low level of conservation (Figure 16). The first polymorphism (*LEC2*_3905801-TTA/) results in a deletion which is also present across the other species for which this part of the sequence was covered (Figure 16). *LEC2*_3905804-C/T, and the second polymorphism, shown after the larger gap in the same image, is however conserved. The third polymorphism (*LEC2*_3905844-ATA/-) is in a conserved locus (*LEC2*_3905849-C/T), while the fourth is in a region with overall low sequence similarity and a locus with a substitution in all but one *L. hirtum* subspecies, and the fifth is in a locus (*LEC2*_3906180-A/G) which is conserved across all the studied species except *L. campestre* (Figure 16).

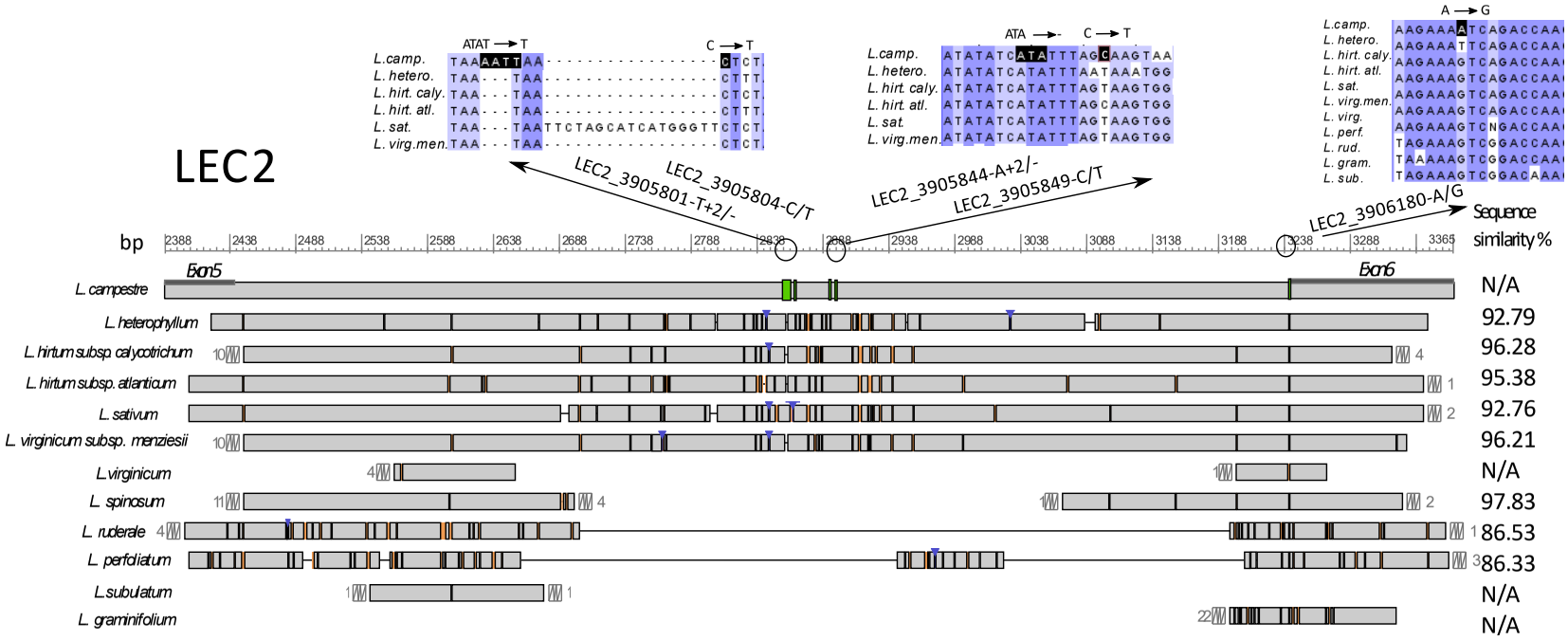


Figure 16. Visualization of the *LEC2* gene alignment in the studied regions between 2388-3365 bp, position relative to CDS from ATG. A band in *L. campestre* indicates where an interesting polymorphism has been found, an orange band in the *Lepidium* accessions indicates substitutions compared to *L. campestre*, a blue band an insertion with an arrow for longer insertions. Sequence similarity is not included for amplicons with <30% coverage of query sequence. Regions around the target polymorphisms (*LEC2*_3905801-T+2/-, *LEC2*_3905804-C/T, *LEC2*_3905844-A+2/-, *LEC2*_3905849-C/T and *LEC2*_3906180-A/G) shown in detail, in blue.

Phylogenetic relationship among the Lepidium accessions

In order to visualize the evolutionary relationship between 15 studied *Lepidium* accessions a phylogenetic tree was generated based on the *FAD2* amplicon (Figure 13). *FAD2* was chosen since it has highest sequence similarity among the four amplified genes (*FAD2*, *TAG1*, *WRI1* and *LEC2*) and therefore assumed to be most conserved. This phylogenetic tree (Figure 17) suggested a close relationship between *L. campestre*, *L. heterophyllum* and *L. hirtum* supported by previous research (Mummenhoff *et al.* 2009, Mummenhoff, Brüggemann & Bowman 2001, Geleta *et al.* 2020), with a larger distance to the other species in the study. *L. perfoliatum* branches out from the other *Lepidium* accessions (Figure 17), despite previous research suggesting a closer relationship (Mummenhoff *et al.* 2009, Mummenhoff, Brüggemann & Bowman 2001). *L. sativum* is placed close to *L. campestre* based on the *FAD2*-sequences (Figure 17) while previous research suggests a more distant relationship (Mummenhoff *et al.* 2009, Mummenhoff, Brüggemann & Bowman 2001).

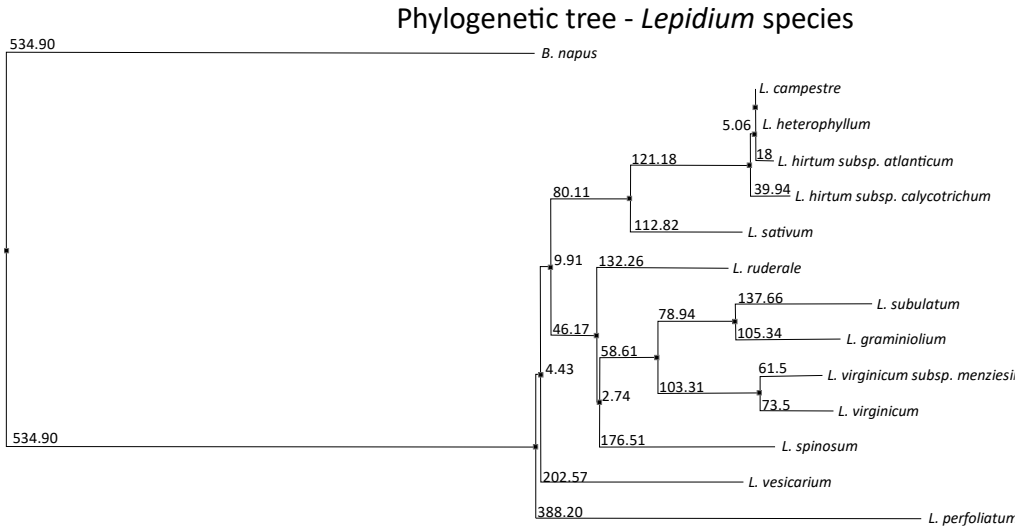


Figure 17. Phylogenetic tree of 15 *Lepidium* accessions based on *FAD2* amplicon (Figure 13) sequences generated with neighbour joining with *B. napus* as an outgroup. Each branch length distance is displayed, with higher numbers indicating greater phylogenetic distance.

3.4.2 Oil content and oil composition analysis in *Lepidium* species

Oil content and oil composition are important traits for domesticating and improving *L. campestre* as a productive oilseed crop. Therefore, the oil content and oil composition of 16 (of which 14 was included in the comparative genomics study) *Lepidium* species were analysed by extracting the seed oil and analysing the fatty acid composition and total fatty acid content on a GC. The highest measured seed oil content was in *L. campestre*, closely followed by *L. ruderale, L. sativum* and *L. virginicum*, all with just above 10% seed oil content (Figure 18). Lowest reported seed oil content was found in the tiny seeds of *L. subulatum* and *L. cardamines*, both below 1% (Figure 18). However, all previously reported values are higher than measured in this study (Nilsson, Johansson & Merker, 1998, Kjaer *et al.* 1954), which is especially noteworthy for *L. graminifolium, L. sativum* and *L. virginicum* where previous research indicates very high oil content. Oil content may vary between individuals, and the low number of seeds may as well be a reason for the difference. The estimated composition of FAs is however closer to previously reported percentages than the total oil content (see Appendix 6, Nilsson, Johansson & Merker 1998, Yaniv *et al.* 1995, Lazzeri *et al.* 2013). The two accessions *L. subulatum* and *L. cardamines* which has very tiny seeds, had a deviating FA composition compared to the rest, with a notably higher proportion of both Myristic acid (14:0), Palmitic acid (16:0) and Behenic acid (22:0) and lower Oleic acid (18:1) and Linolenic (18:3) (Figure 18).

Six accessions had a comparably low level of EA (22:1) – *L. ruderale, L. graminifolium, L. perfoliatum, L. sativum, L. vesicarium* and *L. virginicum subsp. menziesii*, while *L. virginicum* and the majority of the other accessions had EA content comparable to *L. campestre* (Figure 18).

Several accessions had a relatively high level of OA (18:1) including *L. heterophyllum, L. hirtum subsp. atlanticum, L. hirtum subsp. calycotrichum L. spinosum, L. sativum, L. virginicum subsp. menziesii* with around double or more OA compared to *L. campestre* (Figure 18).

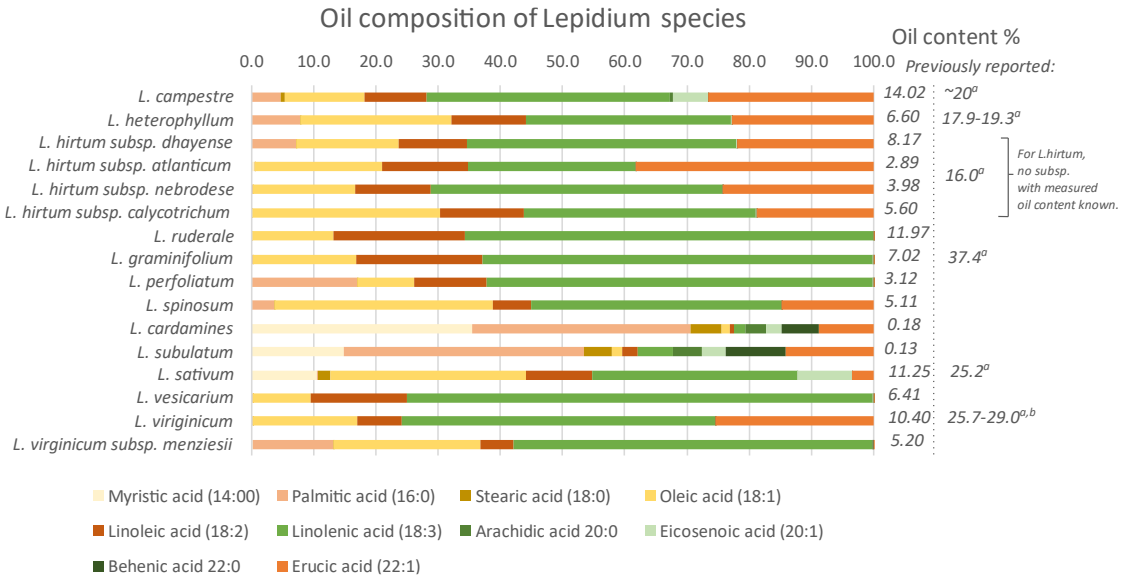


Figure 18. Oil composition in 16 *Lepidium* accessions. n = 10x3 (seeds), except *L. ruderale* and *L. hirtum subsp. atlanticum* with n = 10x2. *Sources for previously reported oil content: [a]Nilsson, Johansson & Merker (1998), [b]Kjaer et al. (1954)*

3.4.3 Lepidium species for future interspecific breeding with *L. campestre*

A literature study was performed to evaluate the potential of other *Lepidium* species for interspecific breeding. From the literature study, it was concluded that there are 14 species of *Lepidium* closely related to *L. campestre* recorded as having the same chromosome number of 2n = 16, or not yet evaluated ploidy level and chromosome number (Table 11). *L. appelianum*, *L. hirtum subsp. atlanticum* and *L. villarsii* have varying chromosome number depending on region (Table 11). Gene bank availability is lacking for many of these species, presenting a challenge if they would be targeted for evaluation in future studies. *L. chalepense* and *L. draba* has been identified as closely related to *L. campestre*, however their chromosome number 2n = 32-64 and 2n = 48-128 respectively will likely be a crossing barrier (Table 11), also supported by a prior study (Reyes 2021). Twenty-two *Lepidium* species were identified as having an ideal chromosome number 2n = 16, but in these cases the phylogenetic distance is likely to become a crossing barrier (Table 11).

Table 11. Table of *Lepidium* species identified with either chromosome number 2n =16 and/or close phylogenetic relationship. Chromosome number & country of origin & alternative name from [a]Warwick & Al-Shehbaz (2006) if not otherwise stated. Relatedness was determined from phylogenetic trees of [b]Mummenhoff *et al.* (2009), [c]Mummenhoff, Brüggemann & Bowman (2001) and [d]Ma *et al.* (2020). Current gene bank availability for distribution is a summary of [e]Genesys (2023) and [f]Kew (2023a) and verified from distributing gene bank webpages. Chromosome number (A) or (B) corresponds to findings in country (A) or (B) in country of origin where chromosome number may be tied to location. Relatedness: Green = Very close within 1-2 nodes (1), Yellow = close, within 3 nodes (2). Brown = Distant, more than 4 nodes (3). Grey = Unknown (4).

Lepidium species	2 n = [a]	Relatedness [d, e, f]	Country of origin [a]	Gene bank – Available for distribution [b, c]	Alternative name [a]
appelianum	16[a], 24 [b]	1	Mongolia, United states & Canada[a], Italy & United states[b]	MSB (no distribution)[c]	*Cardaria pubescens, Hymenophysa pubescens*
heterophyllum	16	1	Czech Rep., Portugal & United Kingdom	DAFF[b], NPGS, IPK[b, c], UPM-BGV [d]	
hirtum	16	1	Morocco & Spain	UPM-BGV[c]	
hirtum subsp. atlanticum	8 (A), 16 (B)	1	Morocco (A), Morocco (B)[a]	NPGS not available[b, c], UPM-BGV [c]	
hirtum subsp. calycotrichum	16	1	Morocco, Spain & Switzerland	NPGS [c]	
hirtum subsp. dhayense	16	1	Morocco	NPGS, UPM-BGV[c]	
hirtum subsp. nebrodense	16	1	Greece & Italy	NPGS[b,c] UPM-BGV [c]	
hirtum subsp. oxyotum	16	1	Greece & France		
hirtum subsp. petrophilum	16	1	Spain		
hirtum subsp. stylatum	16	1	Spain		
perfoliatum	16	2	Canada, Iraq, Unite States, Bulgaria, Czeck/Slovak, Denmark, Iceland, Poland & Russia (USSR)	IPK, NPGS[b, c] MSB, UPM-BGV[c]	
rigidum	uknown	1	Algeria		
villarsii	14 (A), 16 (B)	1	France (A), France (B)	UPM-BGV[c]	
chalepense	48 (A), 80 (B), 128 (C)	1	Iran, Pakistan & Afghanistan (A), Canada & Russia (USSR) (B) Iran (C)	MSB no distribution[c]	
draba	32 (A), 64 (B)	1	Iraq (A), Iran, Bulgaria, Austria, Canada, Czech/Slovak, France, Denmark, Italy, Russia (USSR), Sweden, United Kingdom, United states, Belgium & Germany (B)	LARI[b], AGG[c], NPGS (currently unavailable)[c]	
affghanum	16	3	Afghanistan & Iran		*L. stroganowia*
affine	16	4	Russia (USSR)		
africanum	16 (A), 32 (B)	3	Kenya (A), Australia (B)	MSB[c]	
alluaudii	16	3	Paris, Morocco	UPM-BGV, NPGS[c]	
armoracia	16(A), 32 (B)	3	Sweden (A), Yemen (B)		
aucheri	16	3	Iraq	MSB[c]	
cardamines	16	3	Spain	NPGS[b, c], UPM-BGV[c]	*L. cardamine*
cartilagineum	16 (A), 28 (B), 40 (C)	3	Spain, Iran, Romania (A), Russia(USSR) (B), Finland (C)		
didymum	16 (A), 32 (B)	3	India (A), China, India, Pakistan, United states, Argentina, Spain, Netherlands, Portugal, Denmark,	CIP[b], MSB, AGG[c]	*Coronopus didymus, Senebiera pinnatifida*

			England & Sweden (B)		
graminifolium	16 (A), 48 (B)	3	Denmark (A), Albania, Bulgaria, France, Greece & Spain (B)	NPGS[2,3], UPM-BGV[c]	*L. graminifolium subsp. graminifolium =L. iberis (India)*
jaredii	16	3	United States		
oblongum	16	3	Denmark & Germany	MSB[c]	
obtusum	16	3	Russia(USSR)	MSB[c]	
persicum	16	4	Iran & Afghanistan	UPM-BGV[c]	
persicum subsp. arianum	16+3B	4	Afghanistan		
pseudodidymum	16	4	Chile		
ruderale	16 (A), 32 (B)	3	Iraq, Romania, Russia, Russia (USSR) & Ukraine (A), China, Germany, Belaus, Bulgaria, Czeck/Slovak, Iceland, Poland, Sweden & France (B)	IPK, NPGS[b,c], UPM-BGV, Nordgen[c]	
sativum	16 (A), 24 (B), 32 (C)	3	loc. not given (A), India, Iran, Iraq, Czeck/Slovak, Finland, Germany, Hungary & Poland (B), China (C)	NPGS, PGRC, VIR, IPK, LARI[b], UPM-BGV[c] etc. Prevalent in many seed banks.	
spinosum	16 (A), 24 (B)	3	Spain (A), Spain (B)	LARI[b], NPGS[b,c], MSB, UPM-BGV[c]	*Capsella spinosa,*
subulatum	16	3	Spain	NPGS[b,c], UPM-BGV[c]	
vesicarium	16 (A), 32 (B)	3	Iran, Russia(USSR) (A), Iran (B)	NPGS[b,c], UPM-BGV[c]	
virginicum	16 (A), 32 (B)	3	United States (A), China, Pakistan, Unisted states, Canada, Czeck/Slovak, Poland, Portugal, Denmark (B)	NPGS, CIP[b], UPM-BGV, MSB[c]	*L. menziesii, subsp. Menziesii = subsp. pubescens, L. iberis*

Gene bank name abbreviations: AGG - Australian Grains Genebank, CIP - International Potato Center, IPK - Genebank of IPK Gatersleben, LARI - Lebanese Agricultural Research Institute , MSB - Kew Millennium Seed Bank Kew , NPGS - USDA National Plant Germplasm System, UPM-BGV - The Plant Germplasm Bank of the Polytechnic University of Madrid, VIR - NI Vavilov All-Russian Institute of Plant Genetic Resources.

3.4.4 Glucosinolate content and composition in Lepidium species

Glucosinolate content is an important trait to improve for successful implementation of *L. campestre* as an oilseed crop. A literature study was therefore performed to evaluate GL content in *Lepidium* seeds. The glucosinolate content in seeds is not well documented in the studied *Lepidium* species, with *L. cardamines*, *L. heterophyllum*, *L. hirtum* and *L. spinosum* completely lacking any documented GLs content and composition. Among the species with documented glucosinolate

information, two separate methods for measuring glucosinolate content has been used. More recent references use direct GLs extraction (Arefaine *et al.* 2019, Isoz 2018, Sarakamis & Yanmaz 2011, Đulović *et al.* 2021) while the older study found measures of glucosinolate content as glucose released in sample hydrolysis (Daxenbichler *et al.* 1991). According to the literature findings, species with notably lower glucosinolate than *L. campestre* include *L. graminifolium, L. subulatum* and *L. sativum* (Table 12).

Table 12. Literature study of Lepidum accessions regarding glucosinolate content and composition. The very low GLs content reported in *L. graminifolium* (*) refers to a *L. iberis* accession. Reported values have been converted to µmol/g dry seed weight. Subspecies for *L. hirtum are* not included since no study has evaluated their GL content.

Species	Glucosinolate content	Most abundant glucosinolate
Lepidium campestre	180-360[c], 123-138[e] µmol/g	Sinalbin[c,j]
Lepidium cardamines	*Uknown*	*Uknown*
Lepidium graminifolium	10*-140 µmol/g (d), 123.90±9.98 µmol/g[b]	Glucolepigramin[a], Erucin[d]
Lepidium heterophyllum	*Uknown*	*Uknown*
Lepidium hirtum	*Uknown*	*Uknown*
Lepidium perfoliatum	146 µmol/g [d]	Erucin[g]
Lepidium ruderale	174 µmol/g[d]	Glucotropaeolin[i,j], benzylglucosinolate[h], tropaeolin [d]
Lepidium sativum	127 µmol/g[d], 1.79-4.57µmol/g[g]	glucotropaeolin[d,f,g,j]
Lepidium spinosum	*Uknown*	*Uknown*
Lepidium subulatum	81 µmol/g[d]	glucotropaeolin[d]
Lepidium vesicarium	100 µmol/g[d]	Lepigramin[d]
Lepidium virginicum	105 µmol/g[d], 2021 µmol/g[i]	Sinalbin[d], subsp. menziesii Glucolepidin[j]

[a]Kew (2023b), [b] Đulović et al. (2021), [c]Isoz (2018), [d]Daxenbichler et al. (1991), [e]Arefaine et al. (2019), [f]Gmelin & Virtanen (1959), [g]Sarakamis & Yanmaz (2011), [h]Al-Shehbaz (1986), [i]Kjaer et al. (1954), [j]Lazzeri et al. (2013).

3.4.5 Germination test in *Lepidium* species

A germination test was performed on 25x2 seeds in a greenhouse to evaluate how the *Lepidium* accession seeds germinate to expand knowledge for future use of the accessions in breeding. Seedlings survival in soil was also evaluated on 5x2 seedlings per accession. Germination test clearly shows large differences among the *Lepidium* species in terms of germination success, and ability to survive after transfer to soil.

Among the most successfully germinated species by day 12 after plating were *L. graminifolium*, *L. sativum, L. campestre L. perfoliatum* with 75-

100% germinated seeds (Figure 19). *L. hirtum* had notably varying germination in the subspecies – ranging from 90% in *L. hirtum subsp. dhayense*, to as low as 8% in *L. hirtum subsp. atlanticum* (Figure 19). By day 12, 40-50% of the seeds from *L. vesicarium*, *L. heterophyllum* and *L. cardamines* had germinated, while 20-30% of the *L. spinosum*, *L. subulatum*, *L. virginicum subsp. menziesii*, *L. ruderale* seeds had germinated whereas none of the *L. virginicum* seeds germinated.

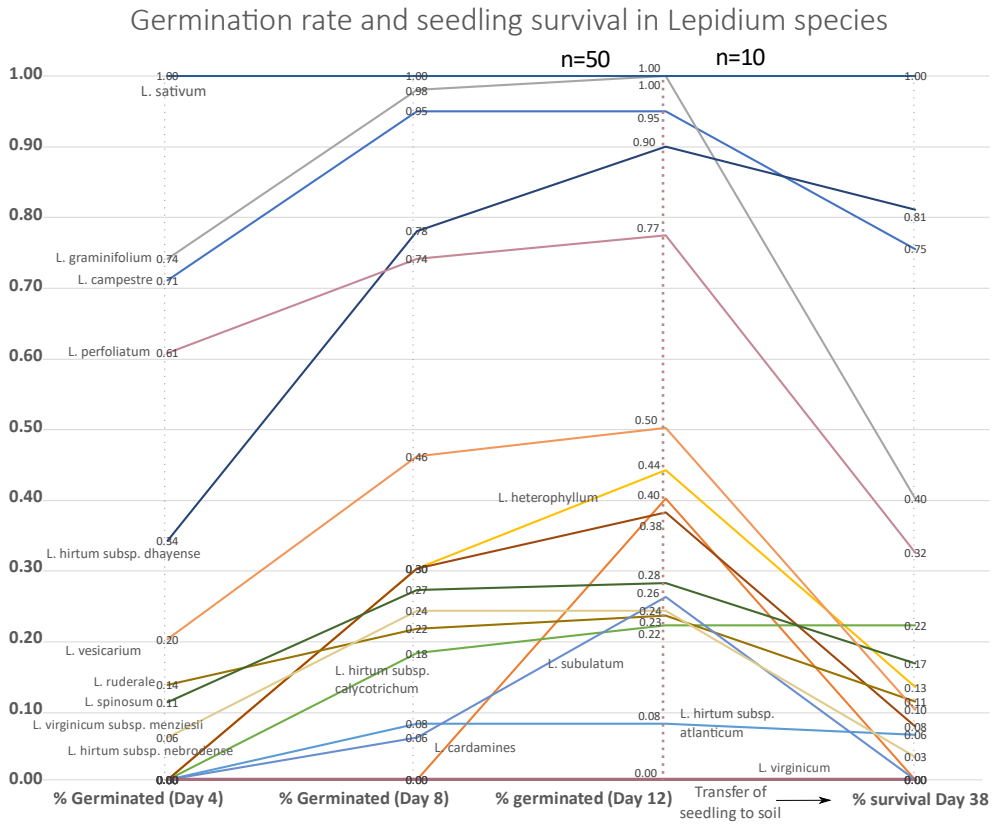


Figure 19. Germination rate (%) of *Lepidium* species at day 4, 8 and 12 after plating seeds (n=50) in a petri dish with sterile water and filter paper and survival rate (% of seeds germinated and survived), of seedlings (n=10) after being placed in soil, and carefully watered three times a week for 38 days.

High germination rate is however not always consistent with high survival rate 26 days after sowing. Although *L. graminifolium* had 100% germination rate (Figure 19), the survival rate of transplanted seedlings was only 40% (Figure 19). Similarly, *L. perfoliatum*, *L. vesicarium*, *L. cardamines* and *L. subulatum* had a germination rate of 77%, 50%, 40% and 26%, respectively at day 12 while only 32%, 10%, 0% and 0%, respectively survived after transplanting to soil. Contrarily, for *L. spinosum* only 28% of the seeds had germinated on day 12, however of those 60% seedlings survived (Figure 19). Most notable in this regard is *L. hirtum subsp. atlanticum* and *subsp. calycotrichum*, where the germination

rate was lowest of all recorded with 8% and 22%, respectively but once relocated in soil, survival was among the highest and almost all seedlings survived transplanting (Figure 19).

Germination rate may be highly dependent on seed age and background factors that were not available at the time of this study and is likely a contributing factor.

The development, morphology, flower induction and seed-setting in *Lepidium* species

Morphology was characterized in 16 accessions of *Lepidium* in 12 species, to expand the knowledge in these species for possible future interspecific hybridization. General morphology was observed and carefully sketched for newly emerged cotyledons, young basal leaves, recently opened flowers and maturing seeds in all species except for *L. cardamines* where no seedlings were obtained and *L. campestre, L. graminifolium, L. vesicarium* and *L subulatum* where flowers were not obtained. Seed characteristics measured included seed weight and area. For flowering it was also observed if the accessions required vernalization prior to flowering, and at what plant height flowering occurred.

Annual accessions with no vernalization requirements identified included *L. ruderale, L. sativum* and *L virginicum* (Table 13).

Table 13. Morphology types (see Figure 20), plant height, phenology, and seed- weight and size for the 16 studied *Lepidium* accessions. Morphology types notes as "N/A" for species where the trait could not be examined due to time constraint. TSW and seed size n = 48 – 1642, plant height at flower induction n = 2-5. Abbreviations: TSW – thousand seed weight.

		L. campestre	*L. heterophyllum*	*L. hirtum subsp. calycotrichum*	*L. Hirtum subsp. atlanticum*	*L. hirtum subsp. dhayense*	*L. hirtum subsp. nebrodese*	*L. perfoliatum*	*L. ruderale*	*L. sativum*	*L. spinosum*	*L. graminifolium*	*L. vesicarium*	*L. virginicum*	*L. virginicum subsp. menziesii*	*L. subulatum*	*L. cardamines*
Morphology types	Cotyledon (Fig. 19A)	1, 3	2	1	2	1	1	3	6	4	6	5	6	2	6	6	N/A
	Basal Leaves (Fig. 19B)	1	1	4	1	1	1	9	7	6	5, 3	2	8	5, 3	5, 3	10	N/A
	Flower (Fig 19C)	N/A	1	1	1	1	1	2	3	7	4	N/A	N/A	5	6	N/A	N/A
	Fruit (Fig 19D)	N/A	1	1	1	1	1	2	3	5	4	N/A	N/A	3	3	N/A	N/A
	Stem (Fig 19E)	N/A	1	1	2	1	1	2	2	2	2	2	N/A	2	2	N/A	N/A
Seed weight and size	Seed weight TSW (g/1000)	2.06	1.08	2.03	1.02	2.09	2.11	0.55	0.20	2.74	0.76	0.24	0.34	0.37	0.48	0.08	0.07
	Seed size (area) mm^2	2.13	1.39	2.22	1.39	2.16	2.42	1.35	0.59	3.16	1.17	0.62	0.68	1.00	1.22	<0.6	<0.6
Flowering	Plant height (cm)	N/A	5	4	17	4	6	26	27.5	18	20	N/A	N/A	32	22.5	N/A	N/A
	Required vernalization?	yes	yes	yes	yes	yes	yes	yes	no	no	yes	yes	yes	no	no	no	N/A

The seeds with the highest seed weight were the semi-domesticated *L. sativum.* Seed weight in most subspecies of *L. hirtum* was found to be comparable to that of *L. campestre* (2.03-2.11 g/1000 seeds), while *L. spinosum, L. heterophyllum*

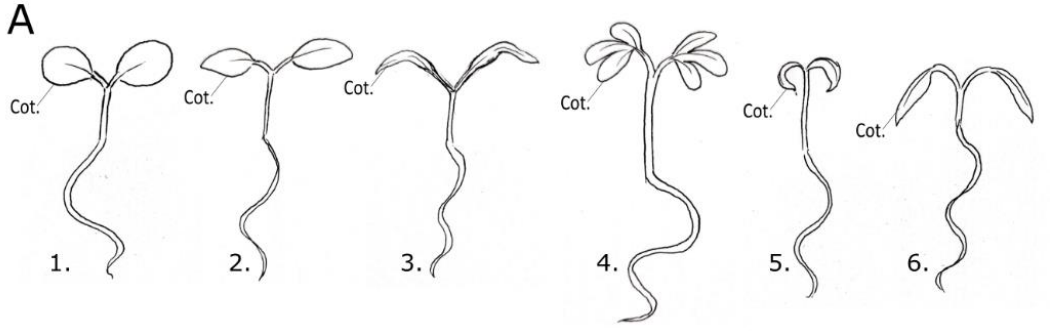
and L. hirtum subsp. atlanticum had less than half of the seed weight compared to *L. campestre* (0.76-1.08 g/1000 seeds) (Table 13). The seed weight of *L. perfoliatum, L. virginicum, L. vesicarium, L. graminifolium* and *L. ruderale* were very low compared to *L. campestre* (0.20-0.48 g/1000 seeds). The two species with generally smallest seeds are *L. subulatum* and *L. cardamines* with less than 0.1 mg/seed (Table 13).

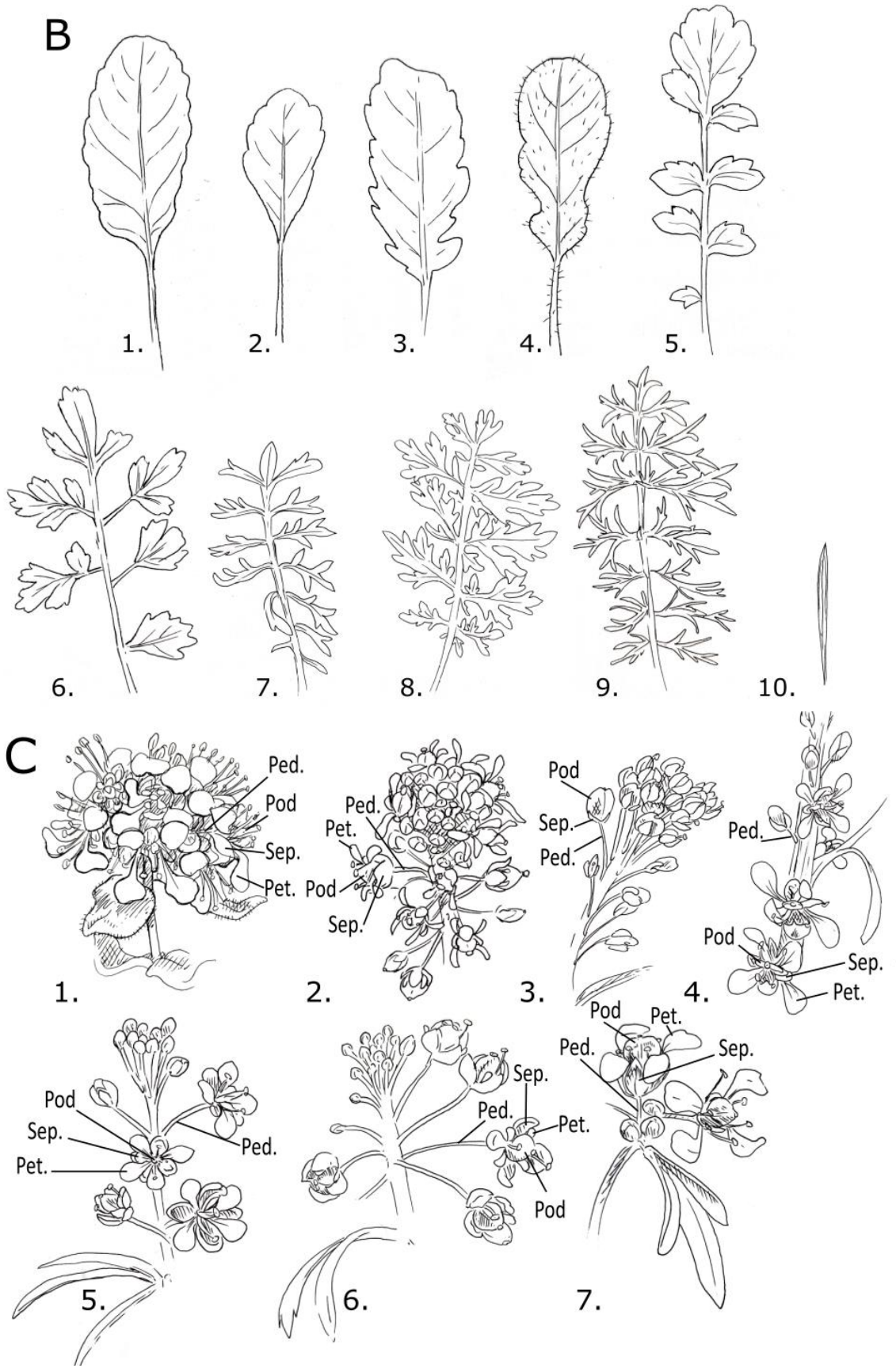
Morphology in cotyledons, leaves and flower was carefully observed and sketched for the studied *Lepidium* accessions, and phenotype was determined (Table 13).

Concerning the cotyledons, *L. sativum* has a distinct three-folded cotyledon leaf (Figure 20A-4) and *L. perfoliatum* have long lanceolate cotyledon leaves (Figure 20A-3). *L. campestre, L. heterophyllum* and *L. hirtum* and *L. virginicum* all have similar round to ovate cotyledon leaves (Figure 20A-1-2).

In terms of basal leaves, species with simple entire, slightly lobed or undulate leaves of similar size include *L. campestre, L. hirtum* and *L. heterophyllum* (Figure 20B-1,3-4), with *L. graminifolium* having smaller simple undulate leaves, elongated in older leaves (Figure 20B-2). *L. spinosum* have entire pinnatifid to pinnatisect leaves (Figure 20B-5), while *L. sativum, L. ruderale, L. vesiarium* and *L. perfoliatum* have entire pinnatisect leaves (Figure 20B-6-9). The most distinguished among the leaves is arguably *L. subulatum* with very small, subulate leaves (Figure 20B-10).

L. heterophyllum and *L. hirtum* inflorescence is initially umbel-like with larger white petals from a short pedicel, on low, not upright standing stems (Figure 20C-1). *L. perfoliatum* is distinct with small flowers with yellow petals on medium-length pedicels (Figure 20C-2). *L. sativum* possesses the largest flowers with large white petals on medium-length pedicels, with a somewhat-umbel-like inflorescence appearance (Figure 20C-7). *L. ruderale* is notably apetalous, with small medium-length pedicels (Figure 20C-3). *L. spinosum* has medium sized flowers with white petals on short thick pedicels along the waxy stem in a spike-like appearance (Figure 20C-4). *L. virginicum* and *L. virginicum subsp. menziesii* has smaller flowers with white petals from a longer pedicel (Figure 20C5-6). *L. virginicum subsp. menziesii* has petals distinctly smaller than the sepals, while *L. virginicum* have slightly larger petals than sepals. L. *virginiucum* accessions were observed to have some pod shattering occur before harvesting.




B
1.
2.
3.
4.
5.
6.
7.
8.
9.
10.
C
Ped.
Pod
Sep.
Pet.
1.
Ped.
Pet.
Pod
Sep.
2.
Pod
Sep.
Ped.
3.
Ped.
Pod
Sep.
Pet.
4.
Pod
Sep.
Pet.
Ped.
5.
Sep.
Ped.
Pet.
Pod
6.
Pod
Pet.
Ped.
Sep.
7.

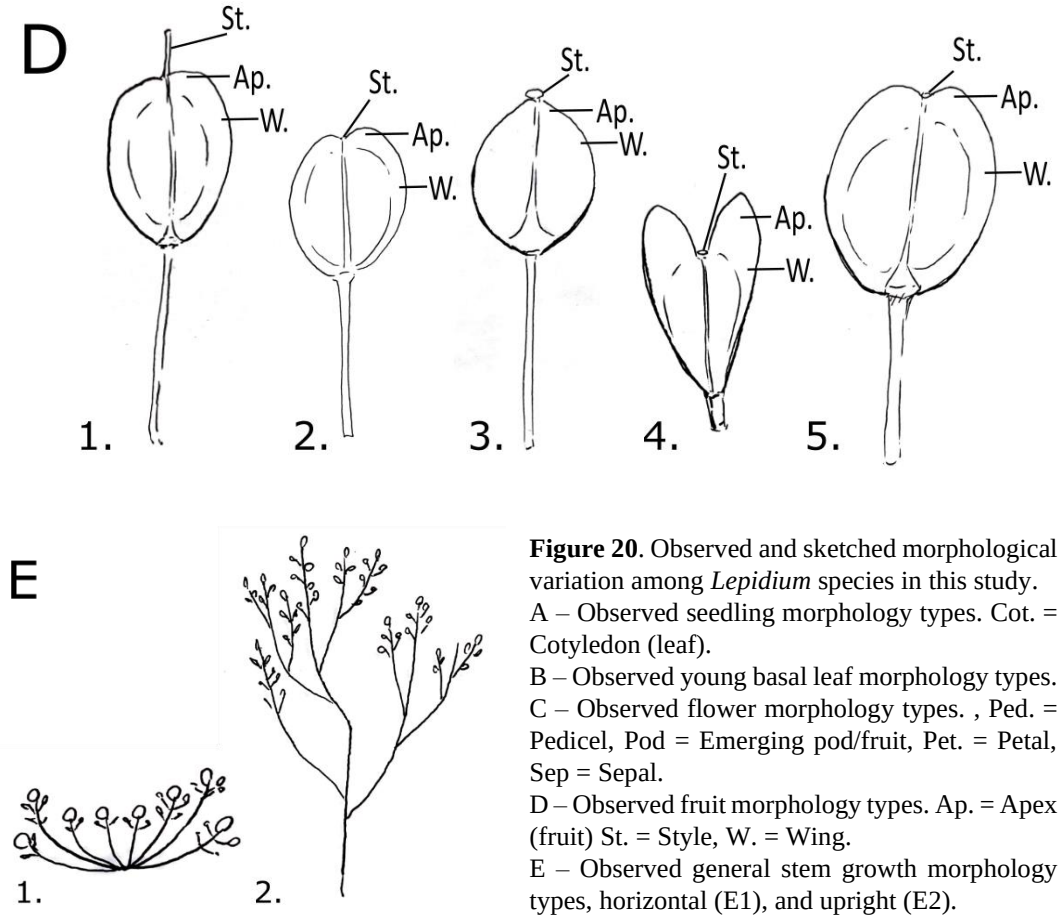


Figure 20. Observed and sketched morphological variation among *Lepidium* species in this study.
A – Observed seedling morphology types. Cot. = Cotyledon (leaf).
B – Observed young basal leaf morphology types.
C – Observed flower morphology types. , Ped. = Pedicel, Pod = Emerging pod/fruit, Pet. = Petal, Sep = Sepal.
D – Observed fruit morphology types. Ap. = Apex (fruit) St. = Style, W. = Wing.
E – Observed general stem growth morphology types, horizontal (E1), and upright (E2).

Regarding fruit morphology, species with oblate to ovate winged fruits with elongated protruding style and none to slightly emarginated apex include *L. campestre, L. hirtum* and *L. heterophyllum*. Whole fruit sparsely to densely covered in hair in *L. hirtum* accessions (Figure 19D-1). Species with winged ovate fruits with slightly to strongly emarginated apex, which include *L. virginicum* including *subsp. menziesii,* and *L. ruderale* (Figure 19C-2), and *L. sativum* (Figure 19C-5). Of these, *L. virginicum* had a tendency for a more strongly emarginated apex, while *L. ruderale* had only a faintly visible mid-wing seed pod line. *L. perfoliatum* was observed to have winged ovate fruit with only a slightly elongated style (Figure 19D-3). *L. spinosum* had winged fruit with distinctly strongly emarginated apex (Figure 19D-4).

In terms of general growth morphology, *L. heterophyllum* and all subspecies of *L. hirtum* except *subsp. atlanticum* was observed to flower at very low height and not grow very tall (Table 13, Figure 19E-1). All other species were observed to have an upright morphology. (Table 13, Figure 19E-2). *L. ruderale* did not grow much taller, but wider, after flowering and seed set.

Summary of evaluation of *Lepidium* species for future interspecific hybridization

A table was created as a summary of all results from this study evaluating *Lepidium* species for interspecific hybridization (Figure 18-19, Table 11-13), for an easier overview (Table 14).

Species with beneficial oil quality (high OA) and short phylogenetic distance to *L. campestre* include *L. heterophyllum* and *L. hirtum* (Table 14). *L. hirtum subsp. calycotrichum* and *dhayense* also have high survival rate of seedlings (Table 14). *L. perfoliatum* have high EA- and low GL-content, and a phylogenetic distance that is suggested by some sources to be close to *L. campestre* (Mummenhoff et al 2009, Mummenhoff, Brüggemann & Bowman 2001), although disputed by the phylogenetic tree constructed by this study (Figure 17)

High OA-, low EA-, high Oil-content and long phylogenetic distance can be found in *L. graminifolium*, *L. sativum* and *L. virginicum*, most with low GL-content as well (Table 14).

Multiple species have either low EA-, low GL-content, or both, but does not possess other beneficial traits, and have a long phylogenetic distance from *L. campestre* (Table 14). These include *L. ruderale*, *L. vesicarium*, *L. cardamines* and *L. subulatum*.. *L. spinosum* have high OA-, low EA-content, however with no other known beneficial traits.

Table 14. Simplified summary of figure 18-19 and Table 11-13, to showcase potential of the 15 *Lepidium* accessions evaluated in this study. Traits designated as medium-level are comparable to those of *L. campestre* while good/very good is better and poor/very poor/abysmal is worse. Medium/poor phylogenetic distance refers to crossing barrier. Column 2n = 16 "yes" refers to the accession having only the ideal chromosome number of 2n = 16, and yes/no that accession are reported to have both 2n =16 and additionally reported chromosome numbers of 2n ≠ 16 which may act as a crossing barrier.

Accession	Germination	Survival	Seed weight	Oil content	OA profile	EA profile	GLS content	Phylogenetic distance	2n = 16
L. heterophyllum	Poor	Poor	Poor	Medium	Very good	Medium	Uknown	Very good	yes
L. hirtum subsp. atlanticum	Very poor	Medium	Poor	Medium	Good	Very poor	Uknown	Good	yes
L. hirtum subsp. calycotrichum	Poor	Very good	Medium	Medium	Very good	Medium	Uknown	Good	yes
L. hirtum subsp. dhayense	Medium	Good	Medium	Medium	Good	Medium	Uknown	Good	yes
L. hirtum subsp. nebrodense	Poor	Poor	Medium	Medium	Good	Medium	Uknown	Good	yes
L. perfoliatum	Medium	Poor	Very poor	Medium	Poor	Very good	Good	Medium	yes
L. graminifolium	Good	Poor	Poor	Very good	Good	Very good	Very good	Poor	yes/no
L. ruderale	Poor	Poor	Very poor	Medium	Medium	Very good	Medium	Poor	yes/no
L. sativum	Good	Very good	Good	Good	Very good	Very good	Very good	Poor	yes/no
L. spinosum	Poor	Poor	Very poor	Medium	Very good	Good	Uknown	Poor	yes/no
L. cardamines	Poor	Abysmal	Abysmal	Very poor	Very poor	Good	Uknown	Poor	yes
L. subulatum	Poor	Abysmal	Abysmal	Very poor	Very poor	Good	Good	Poor	yes
L. vesicarium	Poor	Poor	Very poor	Medium	Poor	Very good	Good	Poor	yes/no
L. virginicum	Very poor	Poor	Very poor	Good	Good	Medium	Good	Poor	yes/no
L. virginicum subsp. menziesii	Poor	Poor	Very poor	Medium	Very good	Very good	Uknown	Poor	yes/no

4. Discussion

4.1 SNP discovery in *L. campestre*

In this study potential genetic polymorphic variation within the available *L. campestre* population, which may be useful as markers for key traits in future breeding efforts were evaluated. A low level of genetic variation was observed in an initial PCA analysis that could explain phenotypic variation (Figure 11). Subsequent unpaired t-test analysis yielded 113 loci with significant association to oil-, OA-, EA-, GL- and Sb-content (Table 7). Of these, 13 marker groups associated with oil-, five associated with EA-, three with OA-, six with GL- content, eight with Sb- and seven with Gla-content (Table 8) were identified as especially interesting. This was due to their genome position in proximity of function-associated genes and/or highly significant associations. Future validation studies are a necessity to validate the identified marker-trait associations. Genotype data from all the 39 analysed accessions showed that there was no single accession with all beneficial markers, which creates a future opportunity for stacking positive alleles. The most promising markers are those associated with oil content, where ten markers potentially can be introduced in accession LEP2020-69 or LEP2020-78, with already relatively high oil content (Appendix 1). LEP2020-78 also has low GL-content. Another option would be to stack three additional positive alleles into one of the 14 accessions having ten markers and high oil content, such as LEP2020-70 or LEP2020-88. The markers associated with glucosinolates are predominately in genes known to regulate oil content or composition. There is some potential to stack molecular marker to increase oil quality. Four beneficial markers have been found in OA, with potential to introduce one in LEP2020-310. Seven markers were found for EA. Since LEP2020-194 and LEP2020-307 already have all EA-markers, a stacking attempt would have to be made on an accession with few markers and already low EA such as LEP2020-198 where four markers could be introduced.

The most promising markers for oil and glucosinolate content are interestingly the same three markers in proximity of the *WRI1* gene (*WRI1*_16471364-C/C+1, C+2, *WRI1*_16472916-G/G+1 and *WRI1*_16473045-C/T). *WRI1* is a master regulator TF primarily known for its role in oil biosynthesis (Figure 2), with few studies to suggest a role in glucosinolate accumulation. It is

however not unheard of for a TF to be implicated in both glucosinolate and seed biosynthesis, as has been shown with MYB76 – proposed to enhance aliphatic GL accumulation, and negatively affect FA-synthesis in *A. thaliana* (Duan *et al.* 2017). Furthermore, there is a logical albeit simple reasoning for why there could be a negative correlation between oil and glucosinolates. Glucose as a finite resource is an important building block in both glucosinolate- and fatty acid-biosynthesis (Figure 2). Glucose is used in the glucolytic part of the fatty acid biosynthesis, in which *WRI1* have previously been shown to have a regulating role (Baud *et al.* 2009). It can however be argued that the extensive research into the function of *WRI1* in *A. thaliana* and *B. napus* would have uncovered this connection. The correlation between oil and glucosinolate content could however be species specific or situation dependent. A significant negative correlation between glucosinolate and oil content has been identified in *B. napus* (Bhardwaj & Hamama, 2000), both significant negative and positive in *B. rapa* (Lionneton *et al.* 2004, Bhardwaj & Hamama, 2000), and significant positive in *E. sativa* (Sukhija *et al.* 1985). Hence, evidence for a specific correlation is inconclusive, even if the role for *WRI1* in glucosinolate biosynthesis is plausible.

Four missense mutations were found in oil content related genes, three in oil composition related genes and six in glucosinolate content related genes (Table 6). Few of the missense mutations could be statistically linked to traits, despite predictions that most are likely detrimental on protein function (Table 6). This may be due to the difficulties working with quantitative traits, where a network of genes influences both oil and glucosinolate content, as is proposed to potentially be the role for *WRI1*.

The genetic analysis and correlation analysis identified two unfavourable patterns, where 1) high oil and high EA is correlated and 2) high GL and high OA (Table 7). This is not ideal for our goals of a high oil-, high oleic acid-, low erucic acid-, and low glucosinolate-content accession. The positive correlation between total oil content and EA levels has some prior support in *B. napus* where a similar pattern has been observed (Azam, Nasim & Iqbal, 2013). However, in *B. napus* interspecific hybridization with *B. rapa* has led to the development of low EA lines with high oil content (Farooq *et al.* 2015).

Only a few of the markers in proximity of target genes were shown to have a significant correlation for oil content or oil composition, and no such markers were found in genes regulating glucosinolate levels. At least for glucosinolates, this plays into a larger pattern of behaviour of wild plants, highlighting the difficulty obtaining useful markers for certain traits. In the wild, glucosinolate content is vital for survival as a defence against herbivory and there will likely be natural selection against mutations with detrimental effect on phenotype as that would result in the plant being eaten.

Some discrepancies were found between the Sanger and NGS sequencing in the polymorphisms. The reason for these is likely partly due to heterozygosity, even though an issue with the sampling and handling is hard to rule out. The three polymorphisms in LEP2020-88 in *FAD2* not detected in targeted NGS is likely due to heterozygosity, supported by the ambiguous nucleotides observed by Sanger sequencing (Table 4). A total of 15 accessions were observed to have SNPs in NGS sequencing but not in Sanger. Possibly reasons for discrepancy however unlikely include somatic mutations with low allele frequency (≥1%) detectable by NGS but not Sanger due to sequencing depth (CD Genomics, 2023). It may of course also simply result from an error in the base calling. Nevertheless, Sanger sequencing results are not considered in the larger statistical analysis portion of this study, which is solely based on the targeted NGS-data.

4.2 Future approaches to improve key traits in *L. campestre*

The genetic diversity within *L. campestre* has proven low, both by this (Figure 11) and by previous studies (Gustafsson *et al.* 2018). A way to overcome limited natural variation in future domestication efforts of *L. campestre* is to turn to genetic modification technologies. However, current regulations at European Union level heavily restrict current use of Genetic Modification (GM) technologies for plant breeders. One possibility around this is to use the not yet restricted albeit time consuming and costly to perform Ethyl Methanesulfonate (EMS) screening. Successful attempts with EMS-screening in *B. napus* to reduce glucosinolates was recently achieved with mutation in *MYB28*-gene (Jhingan, 2023), and increased OA content by mutation in *FAD2* or *ROD2* (Tang *et al.* 2020).

In a project that aims to develop a novel oil- and cover crop in the United States, *Thlaspi arvense* (pennycress), another Brassicaceae, is currently under domestication much like *L. campestre* (Ringling et al, 2019). In the *T. arvense* project, successful efforts to decrease GL- and EA-content have been moving forward with EMS-screening (Ringling *et al.* 2019) and CRISPR-Cas9 (gene editing) (Mcginn *et al.* 2019), rather than traditional MAS. Work to enhance the oil and glucosinolate content and composition with GM-technology in *L. campestre* is also underway (Sandgrind *et al.* 2023), with a working protoplast regeneration protocol (Sandgrind al. 2021). Markers identified for *L. campestre* in this study could after validation, be suitable targets. Additionally, prior oil composition markers identified in *B. napus* for key genes such as *FAD2* (Yang *et al.* 2012), *FAE1* (Wang et al, 2010, Yan *et al.* 2015) and glucosinolates (Hasan *et al.* 2008) can be useful targets. However, traditional knockout mutations in key

genes can have undesirable consequences for the phenotype, which needs to be considered when using this approach.

A study performed by Sandgrind (2022) in *L. campestre* managed to successfully knock out glucosinolate transporters *GTR1* and *GTR2* in *L. campestre.* Despite the single *GTR2* and double *GTR1/GTR2* mutant having almost completely abolished GL-content in their seeds, the mutations also adversely affected growth and seed yield. In another knockout study by Leet et al (2021), a *FAD2* knockout was shown to enhance monosaturated FAs greatly, however with stunted growth in *C. sativa,* and greatly reduced cold hardiness in *A. thaliana* in a study by Miquel *et al.* (1993). Thus, polyunsaturated acids appear necessary for plant membrane maintenance and composition and consequent plant viability at lower temperatures (Wallis & Browse 2002, Caiveau *et al.* 2001), which is important to keep in mind to preserve the beneficial cold hardiness trait in *L. campestre*. Hence, it is critical to consider whether complete knockout or fine-tuning of gene expression is what is required to increase beneficial traits such as high OA content, while still maintaining cold hardiness. This demonstrates complex fine-tuning of genetic regulation required for successful breeding. Methods like CRISPR-Cas9 may still be employed to modify the *GTR1*, *GTR2* and *FAD2* gene expression, for example through substitutions, INDELs of whole codons, or small truncations, which only causes minor effects on the function. Prior domain and motif functionality studies from *L. campestre*, *B. napus* and *A. thaliana* can be of great help when evaluating targets for modification. It may therefore be beneficial for future studies to further characterize additional genes related to oil-, OA-, EA- and GL-content in *L. campestre.*

4.3 Bioinformatics

4.3.1 In Brassicaceae

In the bioinformatic study of four genes of several species in Brassicaceae it was found that, the level of conservation is clearly high in *FAD2*, *FAE1* and *SOT16* while *PEN2* has a very low conservation level.

In the alignment of *FAD2*, only three INDELs are observed. An insertion is found in all three gene copies of *C. sativa* at 29 bp from the start codon, and another in all eight gene copies across three *Brassica* species at position 610 bp. Another interesting observation in the *Brassica* species is the deletion at position 159 bp which is only present in the two pseudogenes, indicating that this region might play an important role in gene function due to the otherwise overall low INDEL-frequency in *FAD2*.

Although conservation is high between Lcamp*FAE1_1* and *FAE1* in other species, the sequence similarity across Lcamp*FAE1_1*, Lcamp*FAE1_2* and

Lcamp*FAE1_3* was found to be low. Similarity was however higher when only comparing Lcamp*FAE1_2* and Lcamp*FAE1_3*. Lcamp*FAE1_3* may also be a pseudogene as it does not possess either a recognizable stop codon or a protein structure recognised by Interpro (Appendix 4). Interestingly, when used as bait query in BLAST searches, Lcamp*FAE1_2* and LcampFAE_3 do show substantially higher sequence similarity to many predicted *FAE1* homologs in the studied Brassicaceae species than they do to Lcamp*FAE1_1*. Research is required to determine if they are true homologues to Lcamp*FAE1_1*. It they are, it is possible that additional gene copies in the *Brassica* and *Arabidopsis* were not found in the BLAST-search due to low sequence similarity and subsequent high E-value.

PEN2 has a very low sequence similarity towards the compared species, which is not surprising given that *PEN2* has a more complex structure than the rest with ten exons of which the first consists of only 10 bp. There are however also gaps present within the exons in the alignment indicating a low conservation level in the coding sequence. One of the more interesting is an insertion in the middle of exon 9, present in all species except *L. campestre* – which may indicate a functional difference. Multiple genus-specific polymorphisms can be seen for *Arabidopsis*, an insertion in intron one, three and nine, and for the *Brassica* species a gap in intron one, four and nine and exon ten. Fewer species were used for comparative genomics of *PEN2* than for other genes in this study, due to the observed low level of conservation. Much like for *FAE1_2* and *FAE1_3*, homologous sequences may have not been included in the BLAST-search results due to low sequence similarity and subsequent E-value. Interestingly, *PEN2* could not be identified in *C. sativa, C. rubella* and *T. arvense* by using *L. campestre PEN2* sequences as bait. Instead of using *L. campestre* as a bait in the BLAST-search, *A. thaliana* could instead be used to find *PEN2* homologous sequences in relatives. A way forward would also be to identify conserved regions within exons of *PEN2* and use them as bait in a BLAST-search instead, to conclusively evaluate whether the *PEN2* gene is present in these species. It may be speculated that low conservation in *PEN2* is due to variation in glucosinolate composition among the Brassicaceae, since sinalbin is only present as a major glucosinolate in *L. campestre* and *S. alba.*

For *SOT16* a high sequence similarity was observed as well as conservation across the studied Brassicaceae. Although *S. alba* is the only of the studied species, besides *L. campestre* to have sinalbin as a major GL component and *SOT16* as partly responsible for sinalbin biosynthesis, the sequence similarity was shown to be lower between these two species compared to others. Some shorter regions across the gene are conserved between *L. campestre* and *S. alba* and no other Brassicaceae species, but there are no major structural differences that can be specifically linked to the species with or without sinalbin as a major GL component.

4.3.2 Sequence analysis of targeted polymorphisms in *FAD2*, *WRI1*, *TAG1* and *LEC2* in *Lepidium*

In the sequencing and subsequent comparative genomics analysis of 14 polymorphic regions of interest in 15 *Lepidium* accessions of 11 different species, it was found that the targeted amplicon in *FAD2* was well conserved, while amplicons in *WRI1*, *TAG1* and *LEC2* are less conserved. This is consistent with expectations as the *FAD2* amplicon resided within an exon, while *WRI1*, *TAG1* and *LEC2* all contained longer stretches of non-coding sequence. However, there were some difficulties achieving successful amplification and some regions could not be properly aligned due to poor sequencing results. What is evident is the close relationship between *L. campestre* and *L. heterophyllum*, with sequence similarity ranging from 97.2-100.0% in all gene amplicons except *LEC2* (at 92.8%). The studied *L. hirtum* subspecies also have very high sequence similarity – in the range between 95.4-99.8%. Lower similarity was observed in amplicons with a majority non-coding sequence.

No INDELs were observed in *FAD2*, as expected from the high sequence similarity in the comparative genomics study of this gene in different Brassicaceae species. The polymorphic locus *FAD2*_16852827-C+2/- in which the alternative allele is resulting in a serine amino acid deletion in *L. campestre* was not recorded in the other species (Figure 12).

The polymorphism *WRI1*_16471364-C/C+1,C+2 resides in a region of *WRI1* which was proven to be difficult to amplify, and because of this little conclusion can be drawn from the sequence aligment (Figure 13A). The three SNPs in amplicon two of *WRI1* (*WRI1*_16472916-G/G+1, *WRI1*_16473045-C/T and *WRI1*_16473352-C/C+7) are all located in highly polymorphic regions (Figure 13B). The alternative alleles *WRI1*_16472916-G/G+1, SNP *WRI1*_16473045-C/T and *WRI1*_16473352-C/C+7 is the cause of minor polymorphisms with little to no predicted impact on gene function (Figure 13).

Conservation for *TAG1* was low, with INDELs present in all *Lepidium* species in amplicon 1, and difficulties amplifying amplicon 2. All three polymorphisms however (*TAG1*_602057-A/G, *TAG1*_604746-C/T & *TAG1*_604920-T/G) appear in well-conserved regions. The alt. allele in *L. campestre* in all three cases is the same nucleotide as the sequenced relatives in those loci.

The conservation level for *LEC2* is very low, especially in the 2688-3188 intronic region, which may be a mutational hotspot. Alignment was not possible due to low sequence similarity in multiple species in this region. In species where alignment was possible, the number of SNP's and INDELs are noticeably higher than flanking intron (2438-2688) and exon regions (3188-3365). The first two polymorphisms of interest occur within a highly polymorphic region (*LEC2*_3905801-TTA/- & *LEC2*_3905804-C/T). The second and third occurs in a

conserved part of this highly polymorphic region (*LEC2*_3905844-ATA/- & *LEC2*_3905849-C/T), while the last occurs in a relatively conserved region flanking the polymorphic region downstream (*LEC2*_3906180-A/G).

In the phylogenetic tree constructed using the *FAD2* amplicon, the close relationship between *L. campestre, L. hirtum* and *L. heterophyllum* was again validated. A close relationship was also suggested between *L. campestre* and *L. sativum* (Figure 16), however contradicted by lower sequence similarity in the other amplicons from *TAG1*, *WRI1* and *LEC2*, and a more distant relationship suggested in previous studies (Mummenhoff *et al.* 2009).

4.4 Evaluation for future interspecific hybridization

Multiple traits were observed, and interesting phenotypes were recorded for several relatives of *L. campestre,* which may be useful in future breeding efforts. Beneficial traits include high oil-, high OA-, low EA- and low GL-content, chromosome number of 2n = 16, close phylogenetic relationship, high seed weight and high germination and survival of seedlings in soil (Table 14).

Researchers domesticating *L. campestre* have performed successful crosses between *L. heterophyllum* and *L. campestre*, and *L. hirtum* and *L. campestre*. Some of the subspecies of *hirtum* included in this study may be beneficial for additional future crosses. Although all subspecies have relatively high levels of OA, *L. hirtum subsp. calycotrichum* is the most promising for crossbreeding with campestre as it also had a very good survival after seedling transplant to soil and a good seed weight comparable to *L. campestre* (Table 14).

There are multiple species with beneficial traits, however with very low predicted crossing success-rate. Embryo rescue protocols may be established with *L. campestre* to increase chance of a viable embryo in these cases. Crossing bridges between *L. campestre* and species of interest may be used as well, they are however predicted to be difficult to achieve due to the long phylogenetic distance between the *L. campestre* clade and the rest of the *Lepidium* tree (Mummenhoff *et al.* 2009). A further challenge to consider for successful breeding is that the chromosome number varies (Table 11) and needs to be verified as 2n = 16 for acquired seeds.

In terms of identifying species that can be used to introgress relevant traits for improving *L. campestre,* L. *graminifolium, L. sativum* and *L. virginicum* all have beneficial high oil-, high OA- and low EA- and low GL-content compared to *L. campestre* (Table 14). *L. virginicum subsp. menziesii*, with previously unknown oil composition, was found to have even higher OA- and lower EA-content than *L. virginicum* and is more ideal as a breeding partner between the two.

Another interesting observation to follow-up with *L. graminifolium* is the low glucosinolate content reported for *L. iberis* (Daxenbichler *et al.* 1991). *Lepidium iberis* samples have however been reported as a synonym for either *L. graminifolium subsp. graminifolium* (Kew 2023b), or *L. virginicum* (Kjaer *et al.* 1954, Bona, 2014, JSTOR 2023). Reports have measured a glucosinolate content significantly higher in both *L. graminifolium* (Đulović *et al.* 2021) and *L. virginicum* (Daxenbichler *et al.* 1991, Kjaer *et al.* 1954) than reported in *L. iberis* (Daxenbichler *et al.* 1991), albeit all three still have a lower content than *L. campestre.* Chromosome number is not known for *L. iberis,* which may be a challenge for future breeding efforts as well.

In terms of species that may be interesting depending on phylogenetic distance, *L. perfoliatum* has been proposed to be more closely related to *L. campestre* than *L. sativum, L. graminifolium* and *L. virginicum,* with ideal chromosome number of 2n =16, seed availability and very interestingly low erucic acid and low glucosinolate content (Table 14). However, the phylogenetic tree constructed in this study suggest a more distant relationship (Figure 16), which is supported by most sequence similarity results from the sequence alignments (Figure 12-15). The number of polymorphisms between *L. perfoliatum* and *L. campestre* is similar to that of distantly related species such as *L. vesicarium, L. virginicum* and *L. ruderale.*

.

In this study, the oil composition was analysed and reported of several *Lepidium* species for the first time. Among those, *L. ruderale* and *L. vesicarium,* were found to have an oil composition similar to *L. graminifolium,* with low level of EA. *L. spinosum* has a composition similar to *L. campestre,* although with higher OA-content. Species with tiny seeds like *L. subulatum* and *L. cardamines,* which were characterized for FA composition for the first time, were shown to have a distinctly different and more diverse seed oil composition compared to the rest. However, not beneficial for the aims of this study, except slightly lower EA and GL levels compared to *L. campestre*. The low cumulative weight of *L. cardamines* and *L. subulatum* seeds may have caused inaccuracies in the oil content and composition results. However, *L. subulatum* and *L. cardamines* also share multiple undesirable traits such as low seed weight, low oil content, and low germination and survival rates, as well as a long phylogenetic distance making them uninteresting from an interspecific hybridization perspective.

Multiple species were identified and proposed for further studies due to close phylogenetic distance with *L. campestre* or ideal chromosome number of 2n = 16 . Relatives of interest found in this study but not evaluated include *L. appelianum* (alt name: *Cardaria pubescens*), *L. villarsii, L. rigidum, L. affine, L. persicum* and *L. pseudodidymum* (Table 11). All except *L. appelianum,* being poorly documented,

and most lacking known seed bank seed availability. *L. appelianum* may also be of interest due to reported pod shattering resistance (Mohammed *et al.* 2019).

It was evident that some of the studied accessions where better suited than others for growing in greenhouse condition. The reason for low germination of *L. subulatum* may be due to its nature as a gypsophite, thriving in chalk and lime-rich soil (Escudero *et al.* 2000), and similar reason may be true for other species with natural habitats in drier more desert like areas such as *L. vesicarium* and *L. graminifolium,* which did germinate but had less survival rate once transferred to soil.

4.5 Conclusion

In conclusion, this study found 113 polymorphisms with significant association for oil-, oleic acid-, erucic acid- and glucosinolate content, of which 27 predicted most promising in breeding efforts of *L. campestre.* Marker validation is now necessary to confirm trait-loci associations. The most promising markers for both oil- and glucosinolate-content are in proximity of master regulator TF *WRI1*. Oil content markers are overall predicted be more interesting due to higher significances, and marker placement in function-associated genes. Stacking markers found in this study is an option for future studies. It is also evident that breeding efforts may benefit from EMS-screening or genetic engineering approaches due to a lack of genetic diversity within *L. campestre*.

There is also potential to use relatives to introduce traits in *L. campestre* through interspecific hybridization. Long phylogenetic distance between the *L. campestre* clade and many studied species with beneficial traits is currently a challenge. In this study six species was found to have potential for future breeding efforts in *L. campestre. L. heterophyllum* and *L. hirtum subsp. calycotrichum* can currently be used in crossing efforts for increased oleic acid-content. *L. graminifolium, L. sativum* and *L. virginicum subsp. menziesii* is not predicted to currently to have a viable offspring in crosses, they are however of interest for targeted embryo rescue protocol development due to their high oil-, high oleic-, low erucic-, and low glucosinolate-content and chromosome number of 2n =16. Low erucic acid content is also found in *L. perfoliatum,* where phylogenetic distance to *L. campestre* needs to be more clearly evaluated. Multiple species have been identified for future trait evaluation due to their short phylogenetic distance to *L. campestre,* most notably *L. appelianum, L. villarsii* and *L. rigidum.*

Lastly, modifications of key genes related to oil or glucosinolate content or quality through EMS-screening or genetic engineering has the potential to enhance desirable traits. Mutants with beneficial phenotype(s) must however

carefully be evaluated, to determine the plants' response to cold temperatures and pest resistance, to avoid losing these valuable traits.

As a final note - *L. campestre* breeding efforts have been an ongoing process, at different capacities, in the last three decades. Current efforts in *L. campestre* breeding focusing on both traditional and genetic engineering approaches are likely to have an effect of the speed of research of the novel oil crop's commercialization within a near future. *L. campestre* has great potential in diversifying commercial oil crops for increased future plant-oil demands, utilizing arable land in the Nordic climate.

References

Al-Shehbaz, I. A. (1986). THE GENERA OF LEPIDIEAE (CRUCIFERAE; BRASSICACEAE) IN THE SOUTHEASTERN UNITED STATES. Journal of the Arnold Arboretum, 67(3), 265–311.

Andersson AAM, Merker A, Nilsson P, Sorensen H, Åman P. (1999). Chemical composition of the potential new oilseed crops Barbarea vulgaris, Barbarea verna and Lepidium campestre. J Sci Food Agric 79: 179-186.

Andersen, T. G., & Halkier, B. A. (2014). Upon bolting the GTR1 and GTR2 transporters mediate transport of glucosinolates to the inflorescence rather than roots. Plant signaling & behavior, 9(1), e27740. https://doi.org/10.4161/psb.27740

Andersen, Tonni & Nour-Eldin, Hussam & Fuller, Victoria & Olsen, Carl & Burow, Meike & Halkier, Barbara. (2013). Integration of Biosynthesis and Long-Distance Transport Establish Organ-Specific Glucosinolate Profiles in Vegetative Arabidopsis. The Plant cell. 25. 10.1105/tpc.113.110890.

Andersen, T. G., Nour-Eldin, H. H., Fuller, V. L., Olsen, C. E., Burow, M., & Halkier, B. A. (2013). Integration of biosynthesis and long-distance transport establish organ-specific glucosinolate profiles in vegetative Arabidopsis. The Plant cell, 25(8), 3133–3145. https://doi.org/10.1105/tpc.113.110890

Arefaine, H., Rydhmer, L., Andersson, R., & Ivarsson, E. (2019). Lepidium cake as a feedstuff for pigs. Livestock Science, 225, 47–52. https://doi.org/10.1016/J.LIVSCI.2019.04.022

Azam, S. M., Nasim, A., & Iqbal, S. (2013). Correlation studies for some agronomic and quality traits in Brassica napus. Agric. Vol. 29, No.4, 2013.

Barthet, V. J., & Daun, J. K. (2011). Seed Morphology, Composition, and Quality. Canola: Chemistry, Production, Processing, and Utilization, 119–162. https://doi.org/10.1016/B978-0-9818936-5-5.50009-7

Baud, S., Wuillã¨Me, S., To, A., Rochat, C. & Lepiniec, L. (2009). Role of WRINKLED1 in the transcriptional regulation of glycolytic and fatty acid

biosynthetic genes in Arabidopsis. The plant journal, 60 (6), s. 933–947. doi:10.1111/j.1365-313x.2009.04011.x

Bednarek, P., Pislewska-Bednarek, M., Svatos, A., Schneider, B., Doubsky, J., Mansurova, M., Humphry, M., Consonni, C., Panstruga, R., Sanchez-Vallet, A., Molina, A., & Schulze-Lefert, P. (2009). A glucosinolate metabolism pathway in living plant cells mediates broad-spectrum antifungal defense. Science (New York, N.Y.), 323(5910), 101–106. https://doi.org/10.1126/science.1163732

Bhardwaj, H. L., & Hamama, A. A. (2000). Oil, erucic acid, and glucosinolate contents in winter hardy rapeseed germplasms. Industrial Crops and Products, 12(1), 33–38. https://doi.org/10.1016/S0926-6690(99)00043-6

Bona, M. (2014). Taxonomic revision of Lepidium L. (Brassicaceae) from Turkey. 44. 31-62.

Bowsher, C., Steer, M., & Tobin, A. (2008). Plant Biochemistry. CRC Press.

Burel, C., Boujard, T., Escaffre, A.-M., Kaushik, S. J., Boeuf, G., Mol, K. A., van der Geyten, S., & Èhn, E. R. K. (2000). Dietary low-glucosinolate rapeseed meal affects thyroid status and nutrient utilization in rainbow trout (Oncorhynchus mykiss). https://doi.org/10.1017/S0007114500000830

Caiveau, O., Fortune, D., Cantrel, C., Zachowski, A. & Moreau, F. (2001). Consequences of ω-6-Oleate Desaturase Deficiency on Lipid Dynamics and Functional Properties of Mitochondrial Membranes ofArabidopsis thaliana. Journal of biological chemistry, 276 (8), s. 5788–5794. doi:10.1074/jbc.m006231200

Cao, S., Zhang, J., Cheng, H. et al. Identification and Evolutionary Analysis of FAD2 Gene Family in Green Plants. Tropical Plant Biol. 14, 239–250 (2021). https://doi.org/10.1007/s12042-020-09276-x

Carlsson, A. S., Yilmaz, J. L., Green, A. G., Stymne, S. & Hofvander, P. (2011). Replacing fossil oil with fresh oil – with what and for what?. European journal of lipid science and technology, 113 (7), s. 812–831. doi:10.1002/ejlt.201100032

CD Genomics. (2023). Targeted Region Sequencing. Retrieved May 15, 2023, from: https://www.cd-genomics.com/targeted-region-sequencing.html

Cheng, C. & Xiao, P. (2022). Evaluation of the correctable decoding sequencing as a new powerful strategy for DNA sequencing. Life science alliance, 5 (8), s. e202101294. doi:10.26508/lsa.202101294

Ching, A., Caldwell, K.S., Jung, M. et al. (2002). SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. BMC Genet. 3, 19. https://doi.org/10.1186/1471-2156-3-19

Chu, D., & Wei, L. (2021). Context-dependent and -independent selection on synonymous mutations revealed by 1,135 genomes of Arabidopsis thaliana. BMC Ecology and Evolution, 21(1), 68. https://doi.org/10.1186/s12862-021-01792-y

Colette Jako, Arvind Kumar, Yangdou Wei, Jitao Zou, Dennis L. Barton, E. Michael Giblin, Patrick S. Covello, David C. Taylor, Seed-Specific Over-Expression of an Arabidopsis cDNA Encoding a Diacylglycerol Acyltransferase Enhances Seed Oil Content and Seed Weight, Plant Physiology, Volume 126, Issue 2, June 2001, Pages 861–874, https://doi.org/10.1104/pp.126.2.861

Cubillos, F. A., Coustham, V., & Loudet, O. (2012). Lessons from eQTL mapping studies: non-coding regions and their role behind natural phenotypic variation in plants. Current Opinion in Plant Biology, 15(2), 192–198. https://doi.org/10.1016/J.PBI.2012.01.005

Dar, A. A., Choudhury, A. R., Kancharla, P. K., & Arumugam, N. (2017). The FAD2 Gene in Plants: Occurrence, Regulation, and Role . In Frontiers in Plant Science (Vol. 8). https://www.frontiersin.org/articles/10.3389/fpls.2017.01789

Daxenbichler, M. E., Spencer, G. F., Carlson, D. G., Rose, G. B., Brinker, A. M., & Powell, R. G. (1991). for Official Use GLUCOSINOLATE COMPOSITION OF SEEDS FROM 297 SPECIES OF WILD PLANTS. 30(8), 2623–2638.

Duan, S., Jin, C., Li, D., Gao, C., Qi, S., Liu, K., Hai, J., Ma, H., & Chen, M. (2017). MYB76 Inhibits Seed Fatty Acid Accumulation in Arabidopsis. Frontiers in plant science, 8, 226. https://doi.org/10.3389/fpls.2017.00226

Đulović, A. et al. (2021). Lepidium graminifolium L.: Glucosinolate Profile and Antiproliferative Potential of Volatile Isolates. Molecules, 26 (17), s. 5183. doi:10.3390/molecules26175183

Eriksson, D. (2009). Towards the Domestication of Lepidium campestre as an Undersown Oilseed Crop. [Doctoral thesis, Swedish Agricultural University, Faculty of Landscape Planning, Horticulture and Agricultural Science Department of Plant Breeding and Biotechnology].

Escudero, A., Iriondo, J. M., Olano, J. M., Rubio, A., & Somolinos, R. C. (2000). Factors Affecting Establishment of a Gypsophyte: The Case of Lepidium subulatum

(Brassicaceae). American Journal of Botany, 87(6), 861–871. https://doi.org/10.2307/2656894

Europeiska kommissionen (EU). (2013). nr 1275/2013 av den 6 december 2013 om ändring av bilaga I till Europaparlamentets och rådets direktiv 2002/32/EG vad gäller gränsvärden för arsenik, kadmium, bly, nitrit, flyktig senapsolja och skadliga botaniska orenheterText av betydelse för EES. https://eur-lex.europa.eu/legal-content/SV/TXT/PDF/?uri=CELEX:32013R1275&from=HR

Fahey, J. W., Zalcmann, A. T., & Talalay, P. (2001). The chemical diversity and distribution of glucosinolates and isothiocyanates among plants. Phytochemistry, 56(1), 5–51. https://doi.org/10.1016/S0031-9422(00)00316-2

Fahlgren, D. (2014). In vitro germination and embryo rescue of Lepidium campestre hybrids. [Bachelor thesis, Swedish Agricultural University, Department of Faculty of Landscape Planning, Horticulture and Agricultural Science].

Farooq, M. A., Ali, Dr. B., Gill, R. A., Islam, F., Cui, P., & Zhou, W. (2015). Breeding Oil Crops for Sustainable Production (pp. 30-45). https://doi.org/10.1016/B978-0-12-801309-0.00002-1

Focks N., Benning C. (1998) wrinkled1: A novel, low-seed-oil mutant of Arabidopsis with a deficiency in the seed-specific regulation of carbohydrate metabolism. Plant Physiol 118: 91-101. doi:10.1104/pp.118.1.91. PubMed: 9733529.

Fuller, D. (2003). Further evidence on the prehistory of sesame. Asian Agri-History. 7. Retrieved April 19, 2023 from: http://www.homepages.ucl.ac.uk/~tcrndfu/articles/Sesame2.pdf

Geleta et al. (2020). High-Density Genetic Linkage Mapping of Lepidium Based on Genotyping-by-Sequencing SNPs and Segregating Contig Tag Haplotypes. Frontiers in Plant Science 11: 448. https://doi.org/10.3389/fpls.2020.00448.

Geleta, M., Zhu, L.-H., Stymne, S., Lehrman, A., & Hansson, S. O. (2013). Domestication of Lepidium campestre as part of mistral biotech, a research pogramme focused on agro-biotechnology for sustainable food. Conference paper, Perennial Crops for Food Security Proceedings of the FAO Expert Workshop.

Genesys PGR. (2023). Retrieved January 25, 2023, from: https://www.genesys-pgr.org/

Ghosh, S., Malhotra, P., Lalitha, P. v., Guha-Mukherjee, S., & Chauhan, V. S. (2002). Novel genetic mapping tools in plants: SNPs and LD-based approaches. Plant Science, 162(3), 329–333. https://doi.org/10.1016/S0168-9452(01)00587-8

Gigolashvili, T., Yatusevich, R., Berger, B., Müller, C. & Flügge, U.-I. (2007). The R2R3-MYB transcription factor HAG1/MYB28 is a regulator of methionine-derived glucosinolate biosynthesis inArabidopsis thaliana. The plant journal, 51 (2), s. 247–261. doi:10.1111/j.1365-313x.2007.03133.x

Gimsing, A. L. & Kirkegaard, J. A. (2009). Glucosinolates and biofumigation: fate of glucosinolates and their hydrolysis products in soil. Phytochemistry reviews, 8 (1), s. 299–310. doi:10.1007/s11101-008-9105-5

Gmelin, R & Virtanen, A. I. (1959). A new type of enzymatic cleavage of mustard oil glucosides. Formation of allylthiocyanate in Thlaspi arvense L. and benzylthiocyanate in Lepidium ruderale L. and Lepidium sativum L. Acta Chem. Scand. 13(7): 1474-1475. doi: 10.3891/acta.chem.scand.13-1474

Gustafsson, C. (2018). Field cress plant in selection based field trials, Alnarp Sweden 2018. Wikimedia Commons. Retrieved May 12, 2023. https://commons.wikimedia.org/wiki/File:20180625_Alnarp4.jpg

Gustafsson, C., Willforss, J., Lopes-Pinto, F., Ortiz, R. & Geleta, M. (2018). Identification of genes regulating traits targeted for domestication of field cress (Lepidium campestre) as a biennial and perennial oilseed crop. BMC genetics, 19 (1). doi:10.1186/s12863-018-0624-9

Hammenhag, C., Varma Saripella, G., Ortiz, R., & Geleta, M. (2020). QTL Mapping for Domestication-Related Characteristics in Field Cress (Lepidium campestre)-A Novel Oil Crop for the Subarctic Region. https://doi.org/10.3390/genes11101223

Harun, S., Abdullah-Zawawi, M.-R., Goh, H.-H., & Mohamed-Hussein, Z.-A. (2020). A Comprehensive Gene Inventory for Glucosinolate Biosynthetic Pathway in Arabidopsis thaliana. Journal of Agricultural and Food Chemistry, 68(28), 7281–7297. https://doi.org/10.1021/acs.jafc.0c01916

Hasan, M. et al. (2008). Association of gene-linked SSR markers to seed glucosinolate content in oilseed rape (Brassica napus ssp. napus). Theoretical and applied genetics, 116 (8), s. 1035–1049. doi:10.1007/s00122-008-0733-3

Hasan, N., Choudhary, S., Naaz, N., Sharma, N. & Laskar, R. A. (2021). Recent advancements in molecular marker-assisted selection and applications in plant breeding programmes. Journal of genetic engineering and biotechnology, 19 (1). doi:10.1186/s43141-021-00231-1

Heather, J. M. & Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. Genomics, 107 (1), s. 1–8. doi:10.1016/j.ygeno.2015.11.003

Hopkins, R. J., Van Dam, N. M. & Van Loon, J. J. A. (2009). Role of Glucosinolates in Insect-Plant Relationships and Multitrophic Interactions. Annual review of entomology, 54 (1), s. 57–83. doi:10.1146/annurev.ento.54.110807.090623

Inkscape Project. (2020). Inkscape. Retrieved from https://inkscape.org

Ivarson, E., Ahlman, A., Lager, I. & Zhu, L.-H. (2016). Significant increase of oleic acid level in the wild species Lepidium campestre through direct gene silencing. Plant cell reports, 35 (10), s. 2055–2063. doi:10.1007/s00299-016-2016-9

Ivarson, E., Leiva-Eriksson, N., Ahlman, A., Kanagarajan, S., Bülow, L., & Zhu, L. H. (2017). Effects of overexpression of WRI1 and hemoglobin genes on the seed oil content of Lepidium campestre. Frontiers in Plant Science, 7, 2032. https://doi.org/10.3389/FPLS.2016.02032/BIBTEX

Isoz, M. (2018). Glucosinolates in Lepidium campestre-Method development and analysis Glukosinolater i Lepidium campestre - Metodutveckling och analysering. [Master thesis, Swedish Agricultural University, Department of Molecular Sciences]. Database: Epsilon (http://stud.epsilon.slu.se)

Jain, J. C., Reed, D. W., GrootWassink, J. W., & Underhill, E. W. (1989). A radioassay of enzymes catalyzing the glucosylation and sulfation steps of glucosinolate biosynthesis in Brassica species. Analytical biochemistry, 178(1), 137–140. https://doi.org/10.1016/0003-2697(89)90369-2

James Jr, D. W., Lim, E., Keller, J., Plooy, I., Ralston, E., & Dooner, H. K. (1995). Directed tagging of the Arabidopsis FATTY ACID ELONGATION1 (FAE1) gene with the maize transposon activator. The Plant Cell, 7(3), 309-319.

Jhingan, S. et al. (2023). Reduced glucosinolate content in oilseed rape (Brassica napus L.) by random mutagenesis of BnMYB28 and BnCYP79F1 genes. Scientific reports, 13 (1). doi:10.1038/s41598-023-28661-6

Jinadasa, B. K. K. K., van Bockstaele, F., Cvejic, J. H., & Simal-Gandara, J. (2022). Current trends and next generation of future edible oils. Future Foods: Global Trends, Opportunities, and Sustainability Challenges, 203–231. https://doi.org/10.1016/B978-0-323-91001-9.00005-0

JMP®, Version 16.2. SAS Institute Inc., Cary, NC, 1989–2023. Retrieved from: https://www.jmp.com/en_in/home.html

Johansson, O. N. et al. (2014). Role of the penetration-resistance genes PEN1, PEN2 and PEN3 in the hypersensitive response and race-specific resistance in Arabidopsis thaliana. The plant journal, 79 (3), s. 466–476. doi:10.1111/tpj.12571

JSTOR. (2023). Lectotype of Lepidium iberis L. [family BRASSICACEAE] on JSTOR. Retrieved March 9, 2023, from https://plants.jstor.org/stable/10.5555/al.ap.specimen.linn-hl824-19

Kerr, R. M., & Dunford, N. (2018). Canola Oil Properties Adding Value to OKLAHOMA. Accessed from: www.fapc.biz.

Kew - Royal Botanical gardens. (2023a) Seed Information Database. Retrieved January 25, 2023, from https://data.kew.org/sid/sidsearch.html

Kew - Royal Botanical gardens. (2023b) Plants of the World online. Retrieved March 3, 2023, from https://powo.science.kew.org/

Kjaer, A., Larsen, I. V. A. N., Tjus, E., & Burris, R. H. (1954). isoThiocyanates. IX. The occurrence of ethyl isothiocyanate in nature. Acta chem. scand, 8(4).

Klein, M. & Papenbrock, J. (2009). Kinetics and substrate specificities of desulfo-glucosinolate sulfotransferases inArabidopsis thaliana. Physiologia plantarum, 135 (2), s. 140–149. doi:10.1111/j.1399-3054.2008.01182.x

Kong Q, Yuan L, Ma W. (2019). WRINKLED1, a "Master Regulator" in Transcriptional Control of Plant Oil Biosynthesis. Plants (Basel). Jul 22;8(7):238. doi: 10.3390/plants8070238. PMID: 31336651; PMCID: PMC6681333.

Kopecký, D., Martín, A. & Smýkal, P. (2022). Interspecific hybridization and plant breeding: From historical retrospective through work of Mendel to current crops. Czech journal of genetics and plant breeding, 58 (No. 3), s. 113–126. doi:10.17221/19/2022-cjgpb

Kruglyak, L. (1999). Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nat Genet 22, 139–144. https://doi.org/10.1038/9642

Kumar, S., Banks, T. W., & Cloutier, S. (2012). SNP Discovery through Next-Generation Sequencing and Its Applications. International Journal of Plant Genomics, 2012, 15. https://doi.org/10.1155/2012/831460

Kumar, N., Chaudhary, A., Singh, D. & Teotia, S. (2020). Transcriptional regulation of seed oil accumulation in Arabidopsis thaliana: role of transcription factors and chromatin remodelers. Journal of plant biochemistry and biotechnology, 29 (4), s. 754–768. doi:10.1007/s13562-020-00616-2

Kumari, J. et al. (2022). Targeted editing of GTR1 and GTR2 homologs directs the development of ideal low-seed and high-leaf glucosinolate oilseed mustard. doi:10.1101/2022.10.03.510517

Kun Wang, John E. Froehlich, Agnieszka Zienkiewicz, Hope Lynn Hersh, Christoph Benning, A Plastid Phosphatidylglycerol Lipase Contributes to the Export of Acyl Groups from Plastids for Seed Oil Biosynthesis, The Plant Cell, Volume 29, Issue 7, July 2017, Pages 1678–1696, https://doi.org/10.1105/tpc.17.00397

Lanza, F. (2011). Olive – A global history. The Edible Series. Reaktion Books.

Lazzeri, L., Malaguti, L. Bagatta, M., D'Avino, L., Ugolini, L., De Nicola, G.R., Casadei, N., Cinti, S., Matteo, R., Iori, R. (2013). Characterization of the main glucosinolate content and fatty acid composition in no food Brassicaceae seeds. Acta Horticulturae. 1005. 331-338. 10.17660/ActaHortic.2013.1005.38.

Lee, DF., Lu, J., Chang, S., Loparo, JJ., Xie, XS. (2016) Mapping DNA polymerase errors by single-molecule sequencing. Nucleic Acids Res. Jul 27;44(13):e118. doi: 10.1093/nar/gkw436. Epub 2016 May 16. PMID: 27185891; PMCID: PMC5291262.

Lee, K. R., Jeon, I., Yu, H., Kim, S. G., Kim, H. S., Ahn, S. J., Lee, J., Lee, S. K., & Kim, H. U. (2021). Increasing Monounsaturated Fatty Acid Contents in Hexaploid Camelina sativa Seed Oil by FAD2 Gene Knockout Using CRISPR-Cas9. Frontiers in plant science, 12, 702930. https://doi.org/10.3389/fpls.2021.702930

Lee JW, Kim IH, Woyengo TA. (2020). Toxicity of Canola-Derived Glucosinolate Degradation Products in Pigs-A Review. Animals (Basel). doi: 10.3390/ani10122337. PMID: 33316893; PMCID: PMC7763053.

Li, X., van Loo, E. N., Gruber, J., Fan, J., Guan, R., Frentzen, M., Stymne, S., & Zhu, L. H. (2012). Development of ultra-high erucic acid oil in the industrial oil crop Crambe abyssinica. Plant Biotechnology Journal, 10(7), 862–870. https://doi.org/10.1111/J.1467-7652.2012.00709.X

Li-Beisson, Y., Shorrosh, B., Beisson, F., Andersson, M. X., Arondel, V., Bates, P. D., Baud, S., Bird, D., DeBono, A., Durrett, T. P., Franke, R. B., Graham, I. A., Katayama, K., Kelly, A. A., Larson, T., Markham, J. E., Miquel, M., Molina, I., Nishida, I., Ohlrogge, J. (2013). Acyl-Lipid Metabolism. The Arabidopsis Book, 11, e0161. https://doi.org/10.1199/TAB.0161

Lionneton, E., Aubert, G., Ochatt, S. & Merah, O. (2004). Genetic analysis of agronomic and quality traits in mustard (Brassica juncea). Theoretical and applied genetics, 109 (4), s. 792–799. doi:10.1007/s00122-004-1682-0

Lipka, V. et al. (2005). Pre- and Postinvasion Defenses Both Contribute to Nonhost Resistance in Arabidopsis. Science, 310 (5751), s. 1180–1183. doi:10.1126/science.1119409

Liu, L. et al. (2012). Comparison of Next-Generation Sequencing Systems. Journal of biomedicine and biotechnology, 2012, s. 1–11. doi:10.1155/2012/251364f

Lebeuf-Taylor, E., Mccloskey, N., Bailey, S. F., Hinz, A. & Kassen, R. (2019). The distribution of fitness effects among synonymous mutations in a gene under directional selection. Elife, 8. doi:10.7554/elife.45952

Ma, J.-B., Wang, T., Wang, X.-R., Qi, D.-S., Li, W.-L., Zhong, , Jiang-Bin, Su, X. U., Zou, D.-L., & Du, Y.-R. (2020). PHYLOGENETIC RELATIONSHIPS OF TWO REPRESENTATIVE LEPIDIUM SPECIES (BRASSICACEAE) IN QINGHAI PROVINCE. Pak. J. Bot, 52(5), 1673–1677. https://doi.org/10.30848/PJB2020-5(19)

Mason, A. S. (2015). SSR Genotyping. I: Methods in Molecular Biology (s. 77–89). Methods in Molecular Biology. doi:10.1007/978-1-4939-1966-6_6

Martinez-Ballesta, M. D. C. & Carvajal, M. (2015). Myrosinase in Brassicaceae: the most important issue for glucosinolate turnover and food quality. Phytochemistry reviews, 14 (6), s. 1045–1051. doi:10.1007/s11101-015-9430-4

Mcginn, M. et al. (2019). Molecular tools enabling pennycress (Thlaspi arvense) as a model plant and oilseed cash cover crop. Plant biotechnology journal, 17 (4), s. 776–788. doi:10.1111/pbi.13014

Meyer, R. S., Duval, A. E. & Jensen, H. R. (2012). Patterns and processes in crop domestication: an historical review and quantitative analysis of 203 global food crops. New phytologist, 196 (1), s. 29–48. doi:10.1111/j.1469-8137.2012.04253.x

Meyer, R. S. & Purugganan, M. D. (2013). Evolution of crop species: genetics of domestication and diversification. Nature reviews genetics, 14 (12), s. 840–852. doi:10.1038/nrg3605

Miquel, M., & Browse, J. (1992). Arabidopsis mutants deficient in polyunsaturated fatty acid synthesis: Biochemical and genetic characterization of a plant oleoyl-

phosphatidylcholine desaturase. Journal of Biological Chemistry, 267(3), 1502–1509. https://doi.org/10.1016/S0021-9258(18)45974-1

Miquel, M., James, D., Dooner, H., & Browse, J. (1993). Arabidopsis requires polyunsaturated lipids for low-temperature survival. Plant Biology, 90, 6208–6212. https://www.pnas.org

Mohammed, S. et al. (2019). Pericarp-mediated chemical dormancy controls the fruit germination of the invasive hoary cress (Lepidium draba), but not of hairy whitetop (Lepidium appelianum). Weed science, 67 (05), s. 560–571. doi:10.1017/wsc.2019.33

Mummenhoff, K., Brüggemann & H., Bowman, J. L. (2001). Chloroplast DNA Phylogeny and Biogeography of Lepidium (Brassicaceae). American Journal of Botany, Vol. 88, No. 11, pp. 2051-2063. Accessed from: https://www.jstor.org/stable/3558431

Mummenhoff, K., Polster, A., Mühlhausen, A. & Theißen, G. (2009). Lepidium as a model system for studying the evolution of fruit development in Brassicaceae. Journal of experimental botany, 60 (5), s. 1503–1513. doi:10.1093/jxb/ern304

NCBI. (2023). ASM975736v2 - Genome - Assembly –Retrieved April 6, 2023, from https://www.ncbi.nlm.nih.gov/assembly/GCA_009757365.2

Nambiar, D. M., Kumari, J., Augustine, R., Kumar, P., Prabodh, |, Bajpai, K., Bisht, N. C., Bajpai, P. K., & Crops, F. (2021). GTR1 and GTR2 transporters differentially regulate tissue-specific glucosinolate contents and defence responses in the oilseed crop Brassica juncea. https://doi.org/10.1111/pce.14072

Nath, U. K., Wilmer, J. A., Wallington, E. J., Becker, H. C., & Möllers, C. (2009). Increasing erucic acid content through combination of endogenous low polyunsaturated fatty acids alleles with Ld-LPAAT + Bn-FAE1 transgenes in rapeseed (Brassica napus L.). TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik, 118(4), 765–773. https://doi.org/10.1007/s00122-008-0936-7

Nilsson, P., Johansson, S. & Merker, A. (1998). Variation in seed oil composition of species from the genera Barbarea and Lepidium. Acta agriculturae scandinavica, section B — soil & plant science, 48 (3), s. 159–164. doi:10.1080/09064719809362493

Nowak, M. A., Boerlijst, M. C., Cooke, J. & Smith, J. M. (1997). Evolution of genetic redundancy. Nature, 388 (6638), s. 167–171. doi:10.1038/40618

Okuley J, Lightner J, Feldmann K, Yadav N, Lark E, Browse J. Arabidopsis FAD2 gene encodes the enzyme that is essential for polyunsaturated lipid synthesis. Plant Cell. 1994 Jan;6(1):147-58. doi: 10.1105/tpc.6.1.147. PMID: 7907506; PMCID: PMC160423.

Ortiz, R., Geleta, M., Gustafsson, C., Lager, I., Hofvander, P., Löfstedt, C., Cahoon, E. B., Minina, E., Bozhkov, P., & Stymne, S. (2020). Oil crops for the future. Current Opinion in Plant Biology, 56, 181–189. https://doi.org/10.1016/J.PBI.2019.12.003

Park, M.-E., Lee, K.-R., Chen, G. Q. & Kim, H. U. (2022). Enhanced production of hydroxy fatty acids in Arabidopsis seed through modification of multiple gene expression. Biotechnology for biofuels and bioproducts, 15 (1). doi:10.1186/s13068-022-02167-1

Pelletier, J. M. et al. (2017). LEC1 sequentially regulates the transcription of genes involved in diverse developmental processes during seed development. Proceedings of the national academy of sciences, 114 (32), s. E6710–E6719. doi:10.1073/pnas.1707957114

Prado, J. R. et al. (2014). Genetically Engineered Crops: From Idea to Product. Annual review of plant biology, 65 (1), s. 769–790. doi:10.1146/annurev-arplant-050213-040039

Rask, L., Andréasson, E., Ekbom, B. et al. Myrosinase: gene family evolution and herbivore defense in Brassicaceae. Plant Mol Biol 42, 93–114 (2000). https://doi.org/10.1023/A:1006380021658

Reintanz, B., Lehnen, M., Reichelt, M., Gershenzon, J., Kowalczyk, M., Sandberg, G., Godde, M., Uhl, R., & Palme, K. (2001). bus , a Bushy Arabidopsis CYP79F1 Knockout Mutant with Abolished Synthesis of Short-Chain Aliphatic Glucosinolates. The Plant Cell, 13, 351–367. www.plantcell.org

Reyes Estévez, E. (2021). In vitro embryo rescue of interspecific hybrids of Lepidium. [Master thesis, Swedish Agricultural University, Faculty of Landscape Architecture, Horticulture and Crop Production Science].Retrieved March 7, 2023, from http://stud.epsilon.slu.se

Ringling K, Chopra R, Anderson N, Marquart L, Marks MD. (2019). Identification and Characterization of Genes Involved in Field Pennycress (Thlaspi arvense L.) Glucosinolate Production. Curr Dev Nutr. Jun 13;3(Suppl 1). doi: 10.1093/cdn/nzz047.OR20-06-19. PMCID: PMC6574296.

Ripa, R.S., Arif, M.R., Islam, M.T. et al. (2020). Embryo rescue response and genetic analyses in interspecific crosses of oilseed Brassica species. In Vitro Cell.Dev.Biol.-Plant 56, 682–693. https://doi.org/10.1007/s11627-020-10116-6

Sandgrind, S. (2022). Genome editing of oilseed species by CRISPR/Cas9 for trait improvement. [Doctoral thesis, Swedish Agricultural University, Faculty of Landscape Architecture, Horticulture and Crop Production Science].

Sandgrind S, Li X, Ivarson E, Ahlman A, Zhu LH. (2021). Establishment of an Efficient Protoplast Regeneration and Transfection Protocol for Field Cress (Lepidium campestre). Front Genome Ed. Nov 16;3:757540. doi: 10.3389/fgeed.2021.757540. PMID: 34870274; PMCID: PMC8635052.

Sandgrind, S., Li, X., Ivarson, E., Wang, E. S., Guan, R., Kanagarajan, S., & Zhu, L. H. (2023). Improved fatty acid composition of field cress (Lepidium campestre) by CRISPR/Cas9-mediated genome editing. Frontiers in plant science, 14, 1076704. https://doi.org/10.3389/fpls.2023.1076704

Sarikamis, G. & Yanmaz, R. (2011). Effects of cultivar and developmental stage on glucosinolates in garden cress (Lepidium Sativum L.). Journal of Medicinal Plants Research. 5.

Shendure, J., Ji, H. Next-generation DNA sequencing. Nat Biotechnol 26, 1135–1145 (2008). https://doi.org/10.1038/nbt1486

Shendure, J. et al. (2017). DNA sequencing at 40: past, present and future. Nature, 550 (7676), s. 345–353. doi:10.1038/nature24286

Schlötterer, C. (2004). The evolution of molecular markers — just a matter of fashion?. Nature reviews genetics, 5 (1), s. 63–69. doi:10.1038/nrg1249

Smith, J. R. (1996). Safflower. The American Oil Chemists Society.

Snell, Per. (2019). Dissecting the gene regulatory networks behind carbon allocation in plants. [Doctoral thesis, Swedish Agricultural University, Faculty of Landscape Architecture, Horticulture and Crop Production Science].

Somerville, C., & Browse, J. (1991). Plant Lipids: Metabolism, Mutants, and Membranes. Science, 252(5002), 80–87. https://doi.org/10.1126/science.252.5002.80

Srinivas, Belide & Petrie, James & Shrestha, Pushkar & Singh, Surinder. (2012). Modification of Seed Oil Composition in Arabidopsis by Artificial microRNA-

Mediated Gene Silencing. Frontiers in plant science. 3. 168. 10.3389/fpls.2012.00168.

Stefansson B. R. & Hougen F. W. 1964. SELECTION OF RAPE PLANTS (BRASSICA NAPUS) WITH SEED OIL PRACTICALLY FREE FROM ERUCIC ACID. Canadian Journal of Plant Science. 44(4): 359-364. https://doi.org/10.4141/cjps64-069

Stoler, N., Nekrutenko, A. (2021). Sequencing error profiles of Illumina sequencing instruments. NAR Genomics and Bioinformatics. Volume 3, Issue 1, March, lqab019. https://doi.org/10.1093/nargab/lqab019

Sukhija, P. S., Loomba, A., Ahuja, K. L., & Munshi, S. K. (1985). Glucosinolates and lipid content in developing and germinating cruciferous seeds. Plant Science, 40(1), 1–6. https://doi.org/10.1016/0168-9452(85)90155-4

Svatoš, A., Nagel, R., Ding, N.-Z., He, M., Qin, C.-X., & Wang, X. (2020). Plant Unsaturated Fatty Acids: Biosynthesis and Regulation. https://doi.org/10.3389/fpls.2020.00390

Tang, S. et al. (2020). Development and screening of EMS mutants with altered seed oil content or fatty acid composition in Brassica napus. The plant journal, 104 (5), s. 1410–1422. doi:10.1111/tpj.15003

TAIR - The Arabidopsis Information Resource (2023). Retrieved March 17, 2023, from https://www.Arabidopsis.org/

Tauger, M. (2010). Agriculture in World History (1st ed.). Taylor and Francis. Retrieved from https://www.perlego.com/book/1608612/agriculture-in-world-history-pdf (Original work published 2010).

Tripathi, M. K., & Mishra, A. S. (2007). Glucosinolates in animal nutrition: A review. Animal Feed Science and Technology, 132(1–2), 1–27. https://doi.org/10.1016/J.ANIFEEDSCI.2006.03.003

Vles, R.O., Bijster, G.M., Timmer, W.G. (1978). Nutritional Evaluation of Low-Erucic-Acid Rapeseed Oils. In: Leonard, B.J. (eds) Toxicological Aspects of Food Safety. Archives of Toxicology, vol 1. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-66896-8_3

Wallis, J. G., & Browse, J. (2002). Mutants of Arabidopsis reveal many roles for membrane lipids. Progress in Lipid Research, 41(3), 254–278. https://doi.org/10.1016/S0163-7827(01)00027-3

Wang, N. et al. (2008). A functional genomics resource for Brassica napus: development of an EMS mutagenized population and discovery of FAE1point mutations by TILLING. New phytologist, 180 (4), s. 751–765. doi:10.1111/j.1469-8137.2008.02619.x

Wang, N., Shi, L., Tian, F. et al. (2010). Assessment of FAE1 polymorphisms in three Brassica species using EcoTILLING and their association with differences in seed erucic acid contents. BMC Plant Biol 10, 137. https://doi.org/10.1186/1471-2229-10-137

Wang, P., Xiong, X., Zhang, X., Wu, G., & Liu, F. (2022). A Review of Erucic Acid Production in Brassicaceae Oilseeds: Progress and Prospects for the Genetic Engineering of High and Low-Erucic Acid Rapeseeds (Brassica napus). Frontiers in Plant Science, 13, 1504. https://doi.org/10.3389/FPLS.2022.899076

Warwick, S.I., Al-Shehbaz, I.A. (2006). Brassicaceae: Chromosome number index and database on CD-Rom. Plant Syst. Evol. 259, 237–248. https://doi.org/10.1007/s00606-006-0421-1

Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. (2009). Jalview Version 2 - A multiple sequence alignment editor and analysis workbench. Bioinformatics 25 1189-1191. (doi:10.1093/bioinformatics/btp033)

Wen, J., Tu, Jx., Li, Zy. et al. (2008). Improving ovary and embryo culture techniques for efficient resynthesis of Brassica napus from reciprocal crosses between yellow-seeded diploids B. rapa and B. oleracea. Euphytica 162, 81–89. https://doi.org/10.1007/s10681-007-9566-4

Wittkop, B., Snowdon, R. J. & Friedt, W. (2009). Status and perspectives of breeding for enhanced yield and quality of oilseed crops for Europe. Euphytica, 170 (1-2), s. 131–140. doi:10.1007/s10681-009-9940-5

Wittstock, U. & Burow, M. (2010). Glucosinolate Breakdown in Arabidopsis: Mechanism, Regulation and Biological Significance. The Arabidopsis book, 8, s. e0134. doi:10.1199/tab.0134

Yan, G., Li, D., Cai, M., Gao, G., Chen, B., Xu, K., Li, J., Li, F., Wang, N., Qiao, J., Li, H., Zhang, T., & Wu, X. (2015). Characterization of FAE1 in the zero erucic acid germplasm of Brassica rapa L. Breeding science, 65(3), 257–264. https://doi.org/10.1270/jsbbs.65.257

Yang, Q., Fan, C., Guo, Z., Qin, J., Wu, J., Li, Q., Fu, T., & Zhou, Y. (2012). Identification of FAD2 and FAD3 genes in Brassica napus genome and development of allele-specific markers for high oleic and low linolenic acid

contents. Theoretical and Applied Genetics, 125(4), 715–729. https://doi.org/10.1007/S00122-012-1863-1/TABLES/4

Yang, J. O., Kim, W. Y., & Bhak, J. (2009). ssSNPTarget: genome-wide splice-site Single Nucleotide Polymorphism database. Human mutation, 30(12), E1010–E1020. https://doi.org/10.1002/humu.21128

Yaniv, Z., Elber, Y., Schafferman, D., Ben-Moshe, E., & Zur, M. (1995). A survey of crucifers native to Israel, as a source of oils. Hulletin dessureshtentius Noticiario de Recursos Fitogenéticos.

Yvert, G., Brem, R., Whittle, J. et al. Trans-acting regulatory variation in Saccharomyces cerevisiae and the role of transcription factors. Nat Genet 35, 57–64 (2003). https://doi.org/10.1038/ng1222

Zohary, D. (2000). Domestication of Plants in the Old World: The Origin and Spread of Cultivated Plants in West Asia, Europe, and the Nile Valley. Oxford University Press

Zhukov, A., & Popov, V. (2022). Synthesis of C20-38 Fatty Acids in Plant Tissues. International journal of molecular sciences, 23(9), 4731. https://doi.org/10.3390/ijms23094731

Popular science summary

Current day fossil fuel usage is a major issue for green gas emissions and its major contribution to global warming. Expanding production of plant-based oils can prove a more sustainable alternative. Rapeseed (*Brassica napus*) is the major oil crop in Sweden today; however, it lacks cold hardiness necessary for Nordic climate. Field cress (*Lepidium campestre*) is a wild plant native to Sweden that is currently under domestication as a future oil crop. It can be grown between other crops in the field and catch excess nutrients in the soil. It is beneficial due to its cold hardiness and high yield with an oil composition suitable for industrial usage. However, multiple traits for improvement have been identified in field cress, including higher oil content and better oil composition. High glucosinolate levels is also an issue with field cress seed-oil. Glucosinolates are compounds present in mainly seeds and leaves, which cause health issues for livestock when they are fed the press cake left after oil extraction. It is an important goal in field cress breeding to decrease the glucosinolates, since it increases sustainability and versatility of the crop when both seeds and press cake have a use.

In this study, we have evaluated 15 key genes important for the regulation of the oil and glucosinolate content and composition in 40 genetically different individuals of field cress. This was done to identify positions in the genetic code that could be linked with beneficial traits, also called “genetic markers”.

The results were that 113 potential markers were found with a significant correlation to oil or glucosinolate content or composition. Of these, ten markers associated with oil content and ten with glucosinolates were identified as especially interesting. This evaluation was based on if markers for a trait was found in genes predicted to influence that trait and had high statistical association. The identified genetic markers can be used for future breeding efforts in improving oil content and glucosinolate levels in field cress.

Fourteen relatives of field cress were evaluated to find beneficial traits that could be introduced in field cress. This would be done through a process called interspecific hybridization, where the two breeding parents are related, but not the same species. These 14 species were evaluated for oil content, oil composition, seed weight, germination of seeds and survival rate of seedlings, plant height on flowering and if induction of flowering required cold treatment. Glucosinolate

content for some species could also be included in evaluation with the help of previous studies.

Multiple species had beneficial traits such as lower glucosinolates, higher seed oil content and beneficial oil composition. The beneficial oil composition, were found in several of the analysed species, including more closely related *L. hirtum* and *L. heterophyllum*. During these studies, it was concluded that even though interesting species for improving field cress was identified, it may prove difficult to cross them with field cress. This is mainly due to the distant relationship between field cress and most of its interesting relatives. Some species can already be crossed with field cress - *L. heterophyllum, L. hirtum* and possibly *L. perfoliatum*. Additional difficulties for further evaluation and attempts at breeding the species include poor seed availability for some of the species and a general lack of information about them.

Proposed future work with *L. campestre* to improve agronomic traits include introducing new variation via mutations, or genetic engineering. Crossing attempts can be made with a subspecies of *L. hirtum*, and *L. heterophyllum*. For species with beneficial traits that cannot currently be crossed, protocols can be developed to increase chance for success.

Acknowledgements

I first and foremost want to thank my truly wonderful supervisor Cecilia, who have shown me respect and great patience throughout this project. Thank you for all the patience, support, encouragement, and problem-solving capabilities but also for allowing me a good amount of freedom with my work. I also want to thank my co-supervisor Mulatu for always having his doors open, ready to discuss various thoughts about my sequences, listen through, and clear any confusion. Niklas and Wes, who should be awarded honorary co-supervisors have been massively helpful and patient with all questions about sequences, PCR, DNA extractions, gels and just about anything. Mirela and Anders for your help and patience with my GC-struggles, and of course, Milla, Ida, Sajeevan & Anna who have always been kind, supportive and very helpful with big and small questions alike. There are so many kind souls in the department who in various ways have been supportive and made me feel welcome, and for this, I am very grateful.

Appendix 1

L. campestre accessions used in this study. Measures in % of total dry weight of seed for oil, % of total oil OA and EA and mg/dry weight in Gla, Sb and GL.

Gla – Glucoallysin, Sb – sinalbin, GL – glucosinolates, Oil – Oil content, OA – Oleic acid, EA – eruic acid.

Phenotype	Sample name	Publ. name	Gla	Sb	GLS	Oil	OA	EA
Oil high	LEP2020-72	BL11	16.41	80.08	96.49	16.72	9.69	28.83
Oil high	LEP2020-80	BL8	22.93	92.74	115.67	17.66	10.35	32.17
Oil high	LEP2020-81	BL10	17.49	108.84	126.33	17.19	10.60	32.47
Oil high	LEP2020-88	LcS1Ger5	18.33	105.66	123.99	18.64	11.60	30.33
Oil high	LEP2020-89	Lc2Fra1	18.37	85.88	104.24	19.90	10.16	32.11
Oil low	LEP2020-198	Lc8Ger1	19.39	92.33	111.73	8.93	13.67	25.74
Oil low	LEP2020-257	LcSar26	31.25	118.71	149.97	9.39	14.90	25.89
Oil low	LEP2020-301	Lc1Cze7	20.62	114.90	145.83	6.52	12.09	28.58
Oil low	LEP2020-303	Lc6Ger1	42.42	141.26	183.68	10.04	17.40	29.61
Oil low	LEP2020-310	Lc3Ger3	36.79	109.23	146.02	5.22	13.35	24.48
OA high	LEP2020-226	LcSse9	17.54	89.03	115.34	12.01	13.55	28.31
OA high	LEP2020-236	LcSse1	24.75	114.68	139.43	14.38	13.51	29.11
OA high	LEP2020-309	Lc3Ger2	32.95	119.43	152.38	10.15	14.86	28.83
OA low	LEP2020-69	BL13	21.05	103.64	124.70	17.30	10.20	31.31
OA low	LEP2020-78	LcS2va10	12.57	62.71	75.28	10.57	8.92	28.04
OA low	LEP2020-234	LcSbj1	22.53	94.42	116.95	14.35	9.42	30.38
EA high	LEP2020-37	LcSgr5	21.65	103.64	125.30	12.68	10.14	32.64
EA high	LEP2020-40	LcS1ho7	23.78	101.90	125.68	15.74	9.93	32.29
EA high	LEP2020-222	Lc1Den8	20.48	99.74	120.22	14.43	11.33	31.73
EA low	LEP2020-21	LcSka6	22.28	100.18	122.46	14.62	10.54	32.50
EA low	LEP2020-194	LcSri7	14.10	88.76	102.86	12.51	13.13	27.23
EA low	LEP2020-302	Lc1Cze5	42.52	135.38	177.90	6.61	11.82	27.42
Gl high	LEP2020-120	LcS2up3	28.88	109.56	138.44	14.54	12.25	29.95
Gl high	LEP2020-126	LcSmo2	26.23	119.67	145.90	10.94	12.36	28.36
Gl high	LEP2020-138	LcSor4	26.39	107.08	133.47	16.38	13.36	29.76
Gl high	LEP2020-262	Lc1Fra2	20.20	130.17	150.37	14.80	11.48	33.09
Gl low	LEP2020-70	BL7	23.52	76.47	99.98	18.30	11.22	30.09
Gl low	LEP2020-92	Lc2Fra4	17.69	80.60	98.29	15.31	9.96	31.63
Gl low	LEP2020-197	LcSsb1b	22.70	91.95	114.65	12.75	13.43	28.41
Sb high	LEP2020-275	Lc1Fra1	18.74	131.42	150.15	13.69	14.04	30.12
Sb high	LEP2020-307	Lc6Ger5	37.48	133.41	170.89	8.75	14.31	28.39
Sb low	LEP2020-23	LcSbj3	23.52	98.41	121.92	12.87	11.91	31.31
Sb low	LEP2020-94	Lc2Gre1	15.37	80.60	95.97	14.48	12.26	29.01
Sb low	LEP2020-192	LcSad7b	17.13	70.07	87.20	13.41	12.16	30.51
Sb low	LEP2020-298	Lc5Ger1	35.27	98.41	133.68	4.76	13.91	27.77
Gla high	LEP2020-3	LcShu14a	30.11	100.33	130.44	11.15	14.71	27.92
Gla high	LEP2020-47	LcSnv3	33.93	116.65	150.59	14.75	14.41	27.50
Gla high	LEP2020-140	LcSar30	29.50	104.96	134.45	13.96	11.63	29.46
Gla low	LEP2020-180	LcSmo1a	9.21	63.54	72.75	14.58	10.90	32.08
Gla low	LEP2020-313	LhSPa1	2.27	109.11	111.38	10.25	12.23	30.72

Appendix 2

Lepidium accessions used in morphology and oil-content study. * LEP88, originally annotated as *L. ruderale*, has been identified as *L. sativum* during this study. Other accessions identified likely to not be ruderale has also been noted with an * and were excluded from all analysis.

Species	Accessions	Acquired from	Country of origin
Lepidium campestre	89	See appendix 1	Sweden
Lepidium campestre	298	See appendix 1	Sweden
Lepidium cardamines	PI 650261	USDA ARS	Spain
Lepidium graminifolium	PI 597855	USDA ARS	Spain
Lepidium heterophyllum	PI 597856	USDA ARS	Spain
Lepidium hirtum subsp. atlanticum	PI 597857	USDA ARS	Morocco
Lepidium hirtum subsp. calycotrichum	PI 597858	USDA ARS	Spain
Lepidium hirtum subsp. dhayense	Ames 21387	USDA ARS	Morocco
Lepidium hirtum subsp. nebrodense	PI 633253	USDA ARS	Italy
Lepidium perfoliatum	Ames 22995	USDA ARS	Germany
Lepidium perfoliatum	PI 650266	USDA ARS	Turkey
Lepidium perfoliatum	PI 633254	USDA ARS	Spain
Lepidium ruderale	PI 597859	USDA ARS	Spain
Lepidium ruderale	PI 633255	USDA ARS	Germany
Lepidium ruderale *	PI 633256	USDA ARS	Poland
Lepidium ruderale	PI 650267	USDA ARS	Hungary
Lepidium ruderale	PI 650268	USDA ARS	Germany
Lepidium ruderale	PI 650269	USDA ARS	Germany
Lepidium ruderale *	PI 650270	USDA ARS	Poland
Lepidium ruderale L.	LEP 98	IPK Gatersleben	Germany
Lepidium ruderale L.	LEP 100	IPK Gatersleben	Germany
Lepidium ruderale L.	LEP 101	IPK Gatersleben	Germany
Lepidium ruderale L.	LEP 81	IPK Gatersleben	Germany
Lepidium ruderale L.	LEP 113	IPK Gatersleben	-
*Lepidium sativum**	LEP 88	IPK Gatersleben	Germany
Lepidium spinosum	PI 633268	USDA ARS	-
Lepidium spinosum	PI 597861	USDA ARS	Turkey
Lepidium subulatum	PI 650272	USDA ARS	Spain
Lepidium vesicarium	PI 650273	USDA ARS	Iran
Lepidium virginicum	Ames 34742	USDA ARS	United states, Arizona
Lepidium virginicum	Ames 34741	USDA ARS	United states, New mexico
Lepidium virginicum	PI 633269	USDA ARS	France
Lepidium virginicum subsp. menziesii	Ames 31357	USDA ARS	United states
Lepidium virginicum subsp. menziesii	PI 633270	USDA ARS	United states

Appendix 3

The sequence of the primers used in this study, and their Tm and target region. Primers names ending with an "F" are sequenced in the 5' → 3' direction, while primer-names ending with an "R" are sequenced in the 3' → 5' direction.

Target	Primer name	Sequence	Tm	Product size
Sanger-study				
FAD2 DNA	LepiFAD2-1a-F	AACGCACTTTCCATTTTTGG	60	**726**
FAD2 DNA	LepiFAD2-1a-R	GGAAGAAATGGCTAGCGAAC	60	
FAD2 DNA	LepiFAD2-2a-F	CCTTCCTCCTCGTCCCTTAC	60.1	**794**
FAD2 DNA	LepiFAD2-2a-R	TTCGCTATTCCTTCTCAATCG	59.4	
FAE1_1 DNA	LepiFAE1_1-1Fa	GCAACTTTGATTGGACGACTG	58.1	**952**
FAE1_1 DNA	LepiFAE1_1-1Ra	CTCACCACTATAAATGCCTTGAGT	58.6	
FAE1_1 DNA	LepiFAE1_1-2Fa	CATTACTCAAGGCATTTATAGTGGTG	59.6	**767**
FAE1_1 DNA	LepiFAE1_1-2Ra	ATTAGGACCGACCGTTTGTG	60.2	
FAE1_1 DNA	LepiFAE1_1-1Fb	CGTCGTCAGTTAACGTCAAACTC	61.1	**574**
FAE1_1 DNA	LepiFAE1_1-1Rb	AACATGCTTGAGTTCACCACA	59.2	
FAE1_1 DNA	LepiFAE1_1-2Fb	CAAGAATACCAAAGTTAACCCTAGAGAGAT	61.8	**996**
FAE1_1 DNA	LepiFAE1_1-2Rb	CGACCGTTTGTGACAGGAGT	61.2	
FAE1_1 DNA	LepiFAE1_1-3Fc	TCTGTCAAAAGACATAACTAATGTTGC	60	**573**
SOT16 DNA	LepiSOT16-1a-F	TCCAGCTTCTTTCATCTTCCA	59.9	**818**
SOT16 DNA	LepiSOT16-1a-R	AACCCATGAAGTCAGCCAAC	60	
SOT16 DNA	LepiSOT16-2a-F	AAGCACGACCCACTGATTTC	59.7	**567**
SOT16 DNA	LepiSOT16-2a-R	GTCCTCTGTTTCCCACCAAA	59.9	
SOT16 DNA	LepiSOT16-3Fa	TCCAGATCGGATTTTGTTCC	59.9	**677**
SOT16 DNA	LepiSOT16-3Ra	TACTTCTTGTTCCGGGCAAT	59.6	
PEN2 DNA	LepiPEN2_1a_F	ACCCACTGAAGATGGCACA	60.1	**640**
PEN2 DNA	LepiPEN2_1a_R	AGATCGCAATTGGATGTTGA	59.1	
PEN2 DNA	LepiPEN2_2a_F	TTAGGCTTTCCATTGCTTGG	60.2	**929**
PEN2 DNA	LepiPEN2_2a_R	CGACAACAAGACAGTTAAAAACG	58.9	
PEN2 DNA	LepiPEN2_3a_F	CGAGCTATGGACTTTATGATTGG	60	**722**
PEN2 DNA	LepiPEN2_3a_R	TTCAGACTCAAAAGGGCTCTAA	58.3	
PEN2 DNA	LepiPEN2_4a_F	GCCAAACAAGGAGGTTCAGA	60.2	**911**
PEN2 DNA	LepiPEN2_4a_R	GGTTCATTTTCAGGCCCATA	59.8	

PEN2 mRNA-characterization				
PEN2 mRNA	LepiPen2mRNA_A1-F	GATGGCACATCTTCAAAGAACA	62.3	**115**
PEN2 mRNA	LepiPen2mRNA_A1-R	TTCACTGCTCCTTCATACTGG	62.6	
PEN2 mRNA	LepiPen2mRNA_A2-F	TCCATCGTTACAAGGAAGATATTAAA	61.5	**100**
PEN2 mRNA	LepiPen2mRNA_A2-R	CTTTTGCCATAAGGTATAACTCG	60.4	
PEN2 mRNA	LepiPen2mRNA_A3-F	CGAGCAGATTATAGATGACTTCAGA	62.8	**287**
PEN2 mRNA	LepiPen2mRNA_A3-R	CGATTTGTCCGTTTTTAATATGAT	59.2	
PEN2 mRNA	LepiPen2mRNA_A4-F	ATGATTGGTTGGCATCATCA	60.9	**235**
PEN2 mRNA	LepiPen2mRNA_A4-R	TGGTTTTCATCCAGTCTACACG	63.3	
PEN2 mRNA	LepiPen2mRNA_A5-F	CGTGTAGACTGGATGAAAACCA	63.3	**275**
PEN2 mRNA	LepiPen2mRNA_A5-R	TCTTCATGGATGGCTTGATG	60.9	
PEN2 mRNA	LepiPen2mRNA_B2-F	CTTACCAGTATGAAGGAGCAGTGA	64.7	**151**
PEN2 mRNA	LepiPen2mRNA_B2-R	CTTTTAATATCTTCCTTGTAACGATGG	61.5	
PEN2 mRNA	LepiPen2mRNA_C2-F	ATTGCTTGGCCACGAGTTAT	63.2	**185**
PEN2 mRNA	LepiPen2mRNA_C2-R	GCTCGCTTAGAAAACCACCA	63.6	
PEN2 mRNA antisense test	LepiPen2mRNA_B1-F	CCCTTTTGCCATCTGTTGAT	61.7	**402**
PEN2 mRNA antisense test	LepiPen2mRNA_B1-R	TGAAGGAGCAGTGAATGTCG	63	
For identifying significant polymorphisms in relatives				
FAD2 CDS	LepiFAD2-1b-F	AGAAACATGGGTGCAGGTG	63.4	**571**
FAD2 CDS	LepiFAD2-1b-R	CGAGGACAAACTGGATGGTT	63.1	
WRI1 Upstream	LepiWRI1-1a-F	TTCCCAAACATAAAACGTACAA	59.7	**973**
WRI1 Upstream	LepiWRI1-1a-R	TTACGCGTGAAGCAGTTGAG	63.7	
WRI1 Intron	LepiWRI1-2b-F	CCGTCAAAGCAGTGGTTTCT	63.8	**927**
WRI1 Intron	LepiWRI1-2b-R	TTCAGAGTATCGGATCCACAAA	62.2	
WRI1 Downstream	LepiWRI1-3a-F	GCCATTATGGAAATGGATCG	59.8	**557**
WRI1 Downstream	LepiWRI1-3a-R	CTGTTCCGTTTTGCAGTTCA	62.2	
TAG1 Intron	LepiTAG1-1a-F	GCAGAGCCATGCTGGATTAT	63.4	**384**
TAG1 Intron	LepiTAG1-1a-R	AAGGCAGCCAAAGGAAAGAT	63.3	
TAG1 Downstream	LepiTAG1-2a-F	AGGTGGGCAACATGATCTTC	63.3	**652**
TAG1 Downstream	LepiTAG1-2a-R	AACACTACGAGCTAGAAAATTCG	61.7	
LEC2 Intron	LepiLEC2-1a-F	AATGGAGCAGAGATGGGAGA	63.5	**978**
LEC2 Intron	LepiLEC2-1a-R	CCAATGAGCATAGCAATGGA	61.2	

Appendix 4

Proposed protein sequence in *L. campestre* for the 15 genes included in this study.

>BUS protein:
MMNLITSLPYPFQILLVFIISMASITLLSQILSRPIKTKDRSRQLPPGPPGWPILGNLPELMRTRPRHKYFQL
AMKNLNSEIGCFNFAGVHAIIINSDEIAREAFKERDADFADRPSHLIGMKTIGDSCKSMGNSPYGEQFQKMKR
VITTDIMSTKSLNMMIASRTIEADNLIAYIHSMYKRSETVDVREFSRVYGYAVTMRLLFGRRHVSKDNAFSDE
GRLGKAEKDHLEAIFNTLNCLPSFSPADYLERWFKGWNIDGQEEMVKQQCGIVRSYNNPIIDERFKLWREKGG
KAAVEDWIDTFITLKDENGKYLITPDEVKAQCEFCIAAIDNPANNMEWTLAEMLKNPEILKKALKELDEVVGR
ERLVQESDIPNLNYLKACCRETFRIHPSAHYVPPHVARRDSTLGGYFIPKGSHIHVGRPAIGRSSKLWKDALV
YKPERHLEGGDTSKEVTLVESEMRLVSFGTGRRGCVGVKVGTIMMVMLLARFLQAFNWKLHPSFGPLSLEEDD
ALLMAKPLLLSVEPRLTPNLYPKFRP-

>*WRI1* protein: - *stop codon not identified
MKKPLSTPTSSSSPSSSSVSSSTTTSSPIQSETLRPKRAKKAKKSSTLSDDKPQSPASTRRSSIYRGVTRHRW
TGRFEAHLWDKSSWNSIQNKKGKQGAYDSEEAAAHTYDLAALKYWGPDTILNFPVETYIKEMEEMQRVTKEEY
LATLRRQSSGFSRGVSKYRGVARHHHNGRWEARIGRVFGNKYLYLGTYNTQEEAAAAYDMAAIEYRGANAVTN
FDISNYIDRLKKKGVFPFPVNQTNHQEASLAEAKQEIETREAKEEPREEVKQQYVEEPQQEQKEEKVEQQEVE
IVGYKEDAVVTCCIDSSAIMEMDRCSDNNELAWNFCMMDSGFAPFLTDQKLSNEKPIEYPELLNELGFEDNID
FMFEEGKNECLSSENLDCCEVVVVGRESPTSSSSPLSCFSTDSASSTTTTSVSCNYSVFRACSLV

>Protein *TAG1*
MAIPDSGGVSTTTENGGGDFADLDRLRRRKSRSDSNGLLSDSPSCGDNLSSDDVGAPNDVRDRIDNVVNDDAQ
GTANLARDNGGDTEIRESGGGRSGGEGRGNVDATFTLRPSVPAHRRVRESPLSSDAIFKQSHAGLFNLCVVVL
VAVNSRLIIENLMKYGWLIRTDFWFSSTSLRDWPLFMCCLSLSIFPLAAFTVEKLALQKCISEPVAIILHIII
TMIEVLYPVYVTLRSCDSAFLSGVTLMLLTCIVWLKLVSYAHTSYDIRSLANSADKGQSSLVLRELKELGIFH
GCSHVMLSGDEMRLQLNSIKHSYTCKSLSSKPRVFGCSQPSYPRSPCIRKGWVARQFAKLIIFTGFMGFIIEQ
QYINPIVRNSKHPLKGDLLYAVERVLKLSVPNLYVWLCMFYCFFHLWLNILAELLCFGDREFYKDWWNAKSVG
DYWRMWNMPVHKWMVRHIYFPCLRHKIPKALAIIIAFLVSAVFHELCIAVPCRLFKLWAFIGIMFQRCLWSLS
QIIYKKGLVQWVGNMIFWFSFCIFGQPMCVLLYYHDLMNRKGSMA-

>*GTR1* protein:
MKSRVILNHRERRDKTNIYTPIDTMERNPLEVETNSYSAVDDGAASNHVISAVDSIDDQQKKLVYRGWKVMPF
IIVNETFEKIGIIGTLSNLLIYLTTVFNLKSYTAATIISAFGGTINFGTFIAAFLCDTYFGRYKTLSVAVIAC
LLGSFVILLTAAVPALHPIACGNKSSCQGPSVGQIMFLMMGLAFLVVGAGGIRPCNLAFGADQFNPKTESGKK
GINSFFNWYFFTFTFAQIISLTLVVYVQSNVSWTIGLTIPVVLMFLACVIFFAGDKLYVKVKASGSPLAGIAH
VIAAAINKRGLKPVKQPWLNLYNHIPPNYANTTLKYTDQFRFLDKAAIMTPEDKLKSDGAASDPWKLCSMQQV
EEVKCIVRVIPIWVASSIYYLAINMQMTYPVFQAVQSDRQLGSGSFRIPGATYVVFLMIGMTIFIIFYDRVFV
PSLRRVTGLDTGITLLQRIGVGIVFATLSLLVSGFIEERRRHIALTKPTLGMARSGEISSMSAFWLIPQLTLA
GIAEAFSAIGQMEFYYKQFPENMRSFAGSIFYVGAGVSNYLSSFLISAVHRTTEHSPTGNWVAEDLNKAKLDY
FYFMLTGVMVVNMAYFLLVAKWYRYKGGNDEDISEIEINEEETKQQQLQDKNSV-

>*GTR2* Protein
MERNTLEVESTDPSSAVYGGSATAVDQEVRDEKKVVYRGWKVMPFIIGNETFEKLGIIGTLSNLLVYLTAVFN
MKSVTAATIINAFSGTINFGTFVAAFLCDTYFGRYKTLSVAVIACFLGSLVILLTAAVPQLHPTPCGSADVCS
GPSGGQVAFLLLGLGFLVVGAGGIRPCNLAFGADQFNPKSESGKRGIDSFFNWYFFTFTFAQILSLTLVVYIQ
SNVSWTIGLTIPAVLMFLACLIFFAGDKLYVKIKASGSPLAGIAQVISVAIKKRGLKPVKQPWLNLYNYYPLN
YANSKLKYTDQFRFLDKAAIMTPEDKLQPDGKPADPWKLCTMQQVEEVKCIVRVLPIWLAASIYYITITQQMT
YPVFQALQSDRRLGSGGFVIPAATYVVFLMTGMTVFIVIYDRVLVPTLKRITGIDTGITLLQRIGTGIFFAIT
SVIVSGFVEERRRTFALTKPTLGMAPRKGEISSMSAMWLIPQLTLAGIAEAFSAIGQMEFYYKQFPENMRSFA
GSVFYVGGGISSYLGSFLIATIHRTTQNSSGGNWLAEDLNKGRLDLFYFMIAGILIVNFIYFLIMSRWYRYKG
SEDEVTAYETNEDLIKQDKNSA-

>*KCS8* protein
MKNLKMFFFKIFFFSLMAGLAMKGSQINFKDLQNFFLHQIQNNLQTIILLLFLGVFVLALYMLTRPKPVYLVD
FSCYLPPSHLKVSIQTLMGHARRAREAGVCWKNKENDYLVDFQEKILERSGLGQETYIPEGLQCFPLQQGMAA
SRKETEEVICGALDNLFRNTGVEPSEIGILVVNSSTFNPTPSLASMIVNKYKFRDNIKSLNLGGMGCSAGVIA
VDTAKGLLQVHRNTYAIVVSTENITQNLYLGKNKSMLVTNCLFRIGGAAVLLSNRSKDRKRAKYELVHTVRIH
TGSDDRSFQCATQEEDEDGIIGVTLTKNLPMVAARTLKINIATLGPLVLPMKEKLAFFLTFIKKKYFNPELKN
YTPDFRLAFEHFCIHAGGRALIDELEKNLKLSPLHVEASRMTLHRFGNTSSSSIWYELAYTEAKGRMKEGDRI
WQIALGSGFKCNSSVWVALRDVKPSANSPWEDCMDRYPVQIDI-

>*LEC1* protein
MERGAPFSHYQLPKSNSGLNLDQHNNPDSTVTSSVVGASNVDTNMTPIGQPQPPCMAREQDQYMPIANVIRIM
RRNLPPHAKISDDAKETIQECVSEYISFVTGEANERCQREQRKTITADDILWAMTKLGFDNYVEPLTVFISRY
REFETDRGCALRGEPTSFKPVYGGNGIGFQGGLPPPGPYGYGMLDPSMVMGGARYFQNGSGQDGSGSGGGSSS
SSMNGMPGFDPYGQYK-

>*LEC2* Protein
MDNFLPFSSNSNANSVQELSMDCNYNRSLFTTIPTYDQAQPHHLLLPYAYPVEQTVPAINPHSLEAFPQIPAM
QTGIEFGSLICNPGLMRQERGGFYDPHMKKMARINRKNAMIRSRNNPSPSSSPNELVDSKKQVMFNIKKTNVP
TADKKDLYRYSSFDNKVWFLFEKLRVLLVKHLKNSDVGTLGRIVLPKREAERNLPELTDKEGILMEMRDVDSI
HSWSFKYKLFWSNNKSRMRVCKKQWSRDGRLLNNIRGRKQEPQQNEGREDESTEVNEMNQYEEIMFDYMIPRD
EDEASIAMLIGNLNDHYPIPNDLMGLTIDLQQHQATSSSSPVNHVHIPRHVGSSDDHVSFNDFTW-

>*MYB28* Protein
MSRKPCCVGEGLKKGAWTTEEDKKLISYIHEHGEGGWRDIPQKAGGLKRCGKSCRLRWTNYLKPEIKRGEFSS
EEEQIIIMLHASRGNNSFRWSVIAKHLPRRTDNEIKNYWNTHLKKRLIDQGIDPVTHKPLASNSKSLVSEDLD
SQDASSSEKQYSRSSSMPSLSKPPVSGSVSEIRNNDGKPVLSDSLSIKKRFKKSSSTSRLLNKVAAKATSIKD
ILSASMEGSLNATTISHARFLNGFSEQVQNEEDSSNASLTNTLSEYDPFSQSSLYPEHEIIATSDLCMDQNYD
FSHFLEGHNFNEETNMNVEYGQDLLMSDMSQEISSTSVDDQDNMVEGWSNYLLDQTDYMYDTDSDSLEKHFI-

>*FAD2* Protein:
MGAGGRMPVPPSSKKTETEALKRVPCEKPPFTLGDLKKAIPPQCFKRSIPRSFSYLISDIIIASCFYYVATNY
FSLLPQSISYLAWPLYWACQGCVLTGIWVIAHECGHHAFSDYQWLDDTVGLIFHSFLLVPYFSWKYSHRRHHS
NTGSLERDEVFVPKQRSAIKWYGKYLNNPLGRVVMLTIQFVLGWPLYLAFNVSGRPYDGFASHFFPNAPIYND
RERLQIYISDAGILAVCYGLYRYAAAQGMASMFCLYGVPLLIVNFFLVLITYLQHTHPSLPHYDSSEWDWLRG
ALATVDRDYGILNKVFHNITDTHVAHHLFSTMPHYNAMEATKAIKPILGEYYQFDGTPWYKAMYREAKECIYV
EPDREDEKKGVYWYNNKL-

>*FAE1_1* Protein:
MTSSVNVKLLYRYVLTNFFNLCLFPLTAFLAGKASKLTANDLYHFYSLLQQNLVTVIILFALIGFGLVLYIVT
RPKPVYLVDYACYLPPPHLKVSISKVIDAFYQIRKADPLRNVACDDPSSLDFLRKIQERSGLGNETYGPEGLV
DVPPRKTFAAAREETEQVINGALENLFKNTKVNPREIGILVVNSSMFNPTPSLSAMVVNTFKLRSNIKSFNLG
GMGCSAGVIAIDLAKDLLQIHKNTYALVVSTENITQGIYSGENRSMMVSNCLFRVGGAAILLSNKPGDRRRSK
YKLAHTVRTHTGADDKSFGCVKQEEDESGKTGVCLSKDITNVAGTTVKKNITTLGPLVLPLSEKFLFFVTFMA
KKLMKDKIKNYYVPDFKLAIDHFCIHAGGRAVIDVLEKNLGLLPIDVEASRSTLHRFGNTSSSSIWYELAYIE
AKGRMEKGNKVWQIALGSGFKCNSAVWVALRNVKPSANSPWEDCIDRYPVELDSDSSKLETPVTNGRS-

>*FAE1_2* Protein:
MANVKLLYHYYLITHFFKLFLLPLMALVAFKASSLNLEDVHNLWFHLQQNFVSLIIPFAVLTFGSAVYFLTRS
RPIYLVDYSCHLPPSHQKVTIQKIIDNVNKNRELNPSMRKLAEDGSLDFFVRVLERSGLGDETCLPDPILNVP
PLNSMAAAREESQQVIFDAIDNLLANTKVNTRDIGIIIVNSSMFNPTPSLSAMVVNKYKLRSTIKSFNLGGMG
CSAGVIAIDLAKDLLQVHKNTYALVVSTENLSRNLYIGDNKSMLVTNCLFRVGGAAILLSNKSGDRRRSKYKL
LHTVRTHTGADDKSFRCVQQEDDDKGKTGVSLTKDITSVASRTITKNIVTLGPLVLPISEKLLFLMTYIRKKI
CDVKIKHYVPDFKRAIDHFCIHAGGRALIDELEKNLGLSPIDVEPSRSTLHRFGNTSSSSIWYELAYTEAKGR
MKKGNKAWQIALGSGFKCNSAVWIALHNVKPSVNSPWEHCVDKYPVNLEF-

>*FAE1_3* * stop codon present prematurely to expected
MANLKVLYHYLITHFFKIFLLPFLTVLAFKAPSLNQEDVQNLWFQHNIVILSIMPLILAFGSFLYFIGRSKPV
YLVDYSCHLPPPHMKVTIPKIIEQITKVRQAHPSMEELADESSLDFLVKIIERSGLGDETYAPEPVTRIPPCQ
SMAAAREETEQVIFDAIDNLLANTKVNPRDIGIIIVNSSMFNPTPSLSAMVVNKYKLRSNIKSFNLGGMGCSA
GVIAIDLAKDLLQVHKNTYALVVSTENLSRNMYIGDNRSMIVTNCLFRVGGAAILLSNKSGDRRRSKYKLLHT
VRTHTGADDKSYFCVQQEDDEKGKTGVSLTKDITSVAARTITKNIATLGPLVLPVSEKCLFFITYMRKKFFDD
KIKHYVPDFKRAIDHFCIHAGGRALIDELEKNLGLSPIDVEPSRSTLHRFGNTSSSSIWYELAYTEAKGRMKK
GHKAWQIALGSGFKCNSAVWIALHNVKPSVNSPWEHCVDKYPVKLDF-FSQX

Appendix 5

Complementary information from the Bioinformatic study of *Brassica*cea (3.1). Seq. length – Sequence length, Acc. No – Accessions number (MCBI), Chr. No. – Chromosome number. (–) indicates species where sequence was not included.

	FAD2			FAE1			PEN2			SOT16		
	Seq. Length	Acc. No	Chr. No	Seq. Length	Acc. No	Chr. No	Seq. Length	Acc. No	Chr. No	Seq. Length	Acc. No	Chr. No
L. campestre	1152	FJ907546.1	unplaced	1521, 1467, 1469	FJ907545.1 unpublished, unpublished	unplaced, unplaced, unplaced	2498	unpublished	unplaced	1014	unpublished	unplaced
A. thaliana	1152	NC_003074.8	3	1521	NM_119617	4	2644	NC_003071.7	2	1017	CP002684.1	1
A. arenosa	1152	LR999453.1	3	1514	LR999457.1	7	2737	LR999454.1	4	1017	LR999452.1	2
A. lyrata	1152	NW_003302553.1	unplaced	1518	GU929425.1	unplaced	3061	NW_003302552.1	4	1017	NW_003302554.1	unplaced
C. sativa	1155, 1158, 930	NC_025685.1, NC_025699.1, NC_025703.1	1, 15, 19	1514, 1514, 1514	GU929420.1, GU929421.1, GU929422.1	10, 11, 12	-	-	-	1014, 1014, 976	NC_025691.1, NC_025693.1, NC_025700.1	7, 9, 16
B. napus	1155, 1155, 1155, 1141 (pseudogene)	JN992606.1, JN992607.1, JN992608.1, JN992609.1	A05, C05, C01, A01	1521, 1521	GU325717.1, GU325719.1	A08, A03	2772, 2821	HG994359.1, HG994368.1	A05, C04	1017, 1020	HG994361.1, HG994370.1	A07, C06
B. rapa	1155, 1141 (pseudogene)	JN859550.1, JN859551.1	A05, A01	1521	GU325723.1	A08	2781	LS974621.2	A05	1017	LR031574.1	A07
B.oleracea	1155, 1155	JN859552.1, JN859553.1	C05, C01	1521	GU325726.1	A03	2825	LR031873.1	C04	1017	LR031880.1	C06
C. rubella	1152	NW_006238921.1	unplaced	1514	NW_006238917.1	unplaced	-	-	-	1014	NW_006238922.1	unplaced
T. arvense	-	-	-	1521	KF664164	5	-	-	-	1014	OU466861.1	5
S. alba	-	-	-	-	-	-	-	-	-	1008, 1014	WIDR02000334.1, WIDR02002676.1	unplaced, unplaced
A. alpina	-	-	-	-	-	-	3001	JF792628.1	4	-	-	-

Appendix 6

Seed oil content (%) and composition (% of total content) found in the 16 studied accessions (light green rows). Also previously reported values of total seed oil content and selected FAs OA & EA (dark green rows). n = 10*3, except *L. ruderale* and *L. hirtum subsp. atlanticum* with n = 10*2. * indicate that *L. hirtum* oil content for specific subsp. not known.

Species	Seed oil content %	Myristic acid % (14:00)	Palmitic acid % (16:0)	Stearic acid % (18:0)	Oleic acid % (18:1)	Linoleic acid % (18:2)	Linolenic acid % (18:3)	Arachidic acid % (20:0)	Eicosenoic acid % (20:1)	Behenic acid % (22:0)	Erucic acid % (22:1)	Reported Seed oil content %	Reported Oleic acid % (18:1)	Reported Erucic acid % (22:1)
L. campestre	14.02	0.1	4.6	0.7	12.9	9.9	39.1	0.6	5.5	0.0	26.7	~20 [a]	15 [a]	22-25 [a]
L. heterophyllum	6.60	0.0	7.8	0.0	24.3	12.1	32.9	0.0	0.1	0.0	22.8	17.9-19.3 [a]	14-20 [a]	23-25 [a]
L. hirtum subsp. dhoyense	8.17	0.0	7.2	0.0	16.5	11.0	43.2	0.0	0.2	0.0	21.9	16.0 [a] *		
L. hirtum subsp. atlanticum	2.89	0.5	0.1	0.0	20.4	13.9	26.9	0.0	0.0	0.0	38.2	16.0 [a] *		
L. hirtum subsp. nebrodese	3.98	0.1	0.1	0.0	16.5	12.2	46.9	0.0	0.1	0.0	24.2	16.0 [a] *		
L. hirtum subsp. colycotrichum	5.60	0.0	0.0	0.0	30.4	13.5	37.2	0.0	0.1	0.0	18.8	16.0 [a] *		
L. ruderale	11.97	0.0	0.0	0.0	13.2	21.1	65.6	0.0	0.0	0.0	0.0			
L. graminifolium	7.02	0.0	0.1	0.0	16.8	20.3	62.8	0.0	0.0	0.0	0.0	37.4 [a]	16.3 [a]	0.3 [a]
L. perfoliatum	3.12	0.0	16.9	0.0	9.1	11.6	62.2	0.0	0.0	0.0	0.0			
L. spinosum	5.11	0.1	3.7	0.0	35.0	6.3	40.3	0.0	0.0	0.0	14.7		18 [b]	5 [b]
L. cardaminea	0.18	35.4	35.1	5.0	1.3	0.7	1.9	3.2	2.5	6.1	8.7			
L. subulatum	0.13	14.8	38.7	4.4	1.6	2.5	5.7	4.7	3.8	9.6	14.2			
L. sativum	11.25	10.5	0.1	2.0	31.5	10.7	33.0	0.0	8.8	0.0	3.4	25.2 [a]	30-35.7 [a,d]	3.3 [a]
L. v esicarium	6.41	0.2	0.0	0.0	9.4	15.4	74.9	0.0	0.0	0.0	0.0			
L. virginicum	10.4	0.1	0.0	0.0	16.8	7.2	50.4	0.0	0.0	0.0	25.4	25.7-29 [a,c]	6.6 [a]	17.9 [a]
L. virginicum subsp. menziesii	5.2	0.2	13.0	0.0	23.5	5.4	57.8	0.0	0.0	0.0	0.0			

[a]*Nilsson, Johansson & Merker (1998),* [b]*Yaniv et al. (1995),* [c]*Kjaer et al. (1954),* [d]*Lazzeri et al. (2013)*

Publishing and archiving

Approved students' theses at SLU are published electronically. As a student, you have the copyright to your own work and need to approve the electronic publishing. If you check the box for **YES**, the full text (pdf file) and metadata will be visible and searchable online. If you check the box for **NO**, only the metadata and the abstract will be visible and searchable online. Nevertheless, when the document is uploaded it will still be archived as a digital file. If you are more than one author, the checked box will be applied to all authors. You will find a link to SLU's publishing agreement here:

- https://libanswers.slu.se/en/faq/228318.

☒ YES, I/we hereby give permission to publish the present thesis in accordance with the SLU agreement regarding the transfer of the right to publish a work.

☐ NO, I/we do not give permission to publish the present work. The work will still be archived and its metadata and abstract will be visible and searchable.