

FROM SPECIES TO TRAIT EVOLUTION  
IN AETHIONEMA (BRASSICACEAE)



FROM SPECIES TO TRAIT EVOLUTION IN AETHIONEMA (BRASSICACEAE)

SETAREH MOHAMMADIN

2017

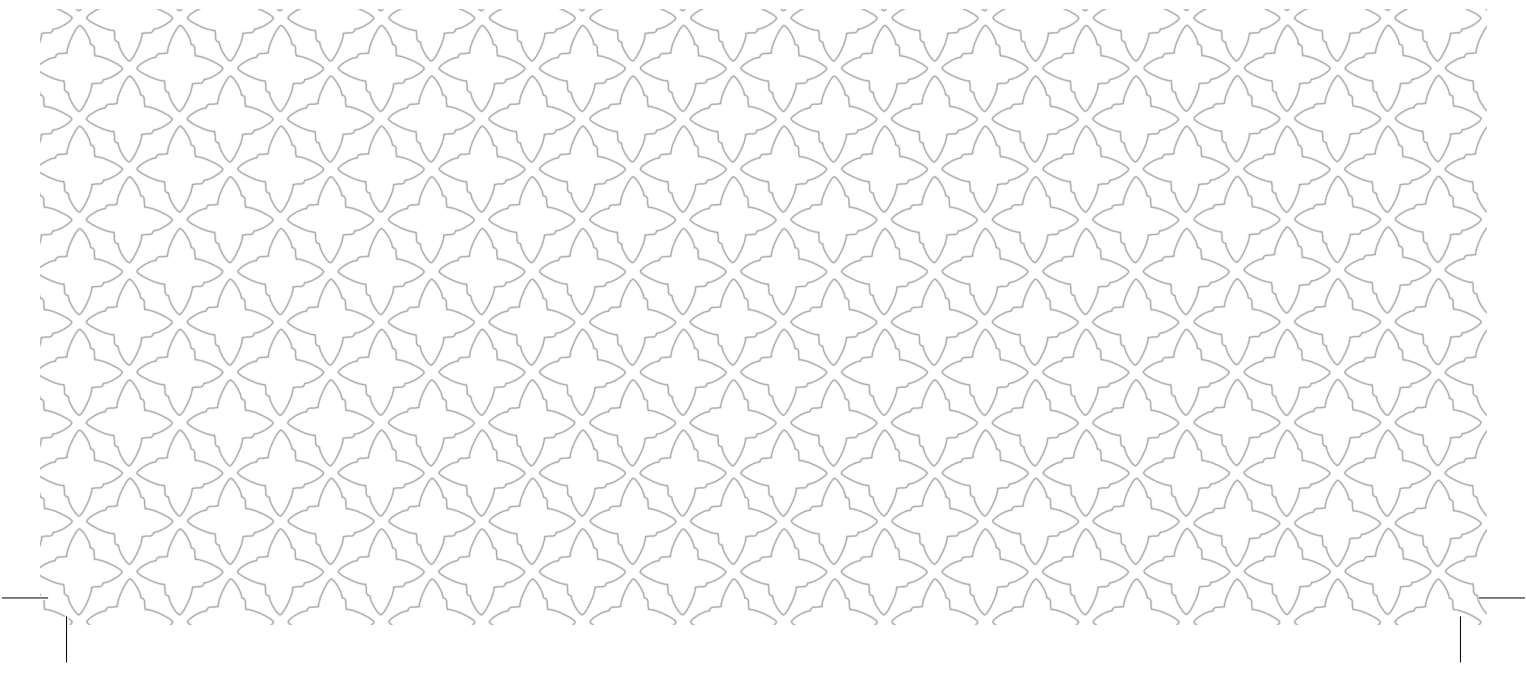
SETAREH MOHAMMADIN



## Propositions

1. Combining multiple techniques makes it possible to analyse the evolution of a species and/or clade from different angles.  
(This thesis)
2. Do not ignore the genomic position, even if the sequence does not match.  
(This thesis)
3. The observation of gravitational waves in 2016 shows that sometimes science can only progress by going back to the basics.
4. Machine Learning will help us to integrate natural sciences together.
5. If someone questions your ideas consider that 'if's' and 'buts' were clusters of nuts, we'd all have a bowl of granola. (Adjusted from Stephen Colbert, The Colbert Report, in an interview with Katheryn Bigelow, January 22, 2013)
6. Successful career-moms will not only pave the way for other women, but also for career-dads.

Propositions belonging to the thesis entitled:  
"From species to trait evolution in *Aethionema* (Brassicaceae)"  
Setareh Mohammadin  
Wagenigen, 11 April, 2017



From species to trait evolution  
in *Aethionema* (Brassicaceae)

Setareh Mohammadin

## **Thesis committee**

### **Promotor**

Prof. Dr M.E. Schranz  
Professor of Biosystematics  
Wageningen University & Research

### **Other members**

Prof. Dr B.J. Zwaan, Wageningen University & Research  
Dr V.I.D. Ros, Wageningen University & Research  
Dr F.P. Lens, Naturalis Biodiversity Center, Leiden  
Dr M. Burow, University of Copenhagen, Denmark

This research was conducted under the auspices of the Graduate School of Experimental Plant Sciences.

# From species to trait evolution in *Aethionema* (Brassicaceae)

Setareh Mohammadin

## **Thesis**

submitted in fulfilment of the requirements for the degree of doctor  
at Wageningen University  
by the authority of the Rector Magnificus,  
Prof. Dr A.P.J. Mol,  
in the presence of the  
Thesis Committee appointed by the Academic Board  
to be defended in public  
on Tuesday 11 April 2017  
at 11 a.m. in the Aula.

Setareh Mohammadin  
From species to trait evolution in *Aethionema* (Brassicaceae),  
126 pages.

PhD thesis, Wageningen University, Wageningen, the Netherlands  
(2017)  
With references, with summary in English

ISBN 978-94-6343-138-5  
DOI 10.18174/409855

## CONTENTS

Abbreviations	7
<b>Chapter 1</b>	<b>9</b>
General introduction	
<hr/>	
<b>Chapter 2</b>	<b>15</b>
Positionally-conserved but sequence-diverged: Identification of Long non-coding RNAs in the Brassicaceae and Cleomaceae	
<hr/>	
<b>Chapter 3</b>	<b>39</b>
Anatolian origins and diversification of <i>Aethionema</i> , the sister lineage of the core Brassicaceae	
<hr/>	
<b>Chapter 4</b>	<b>59</b>
Genome-wide nucleotide diversity and associations with geography, ploidy level and glucosinolate profiles in <i>Aethionema arabicum</i> (Brassicaceae)	
<hr/>	
<b>Chapter 5</b>	<b>79</b>
Major multi-trait quantitative trait locus controls glucosinolate content across developmental stages of <i>Aethionema arabicum</i> (Brassicaceae)	
<hr/>	
<b>Chapter 6</b>	<b>97</b>
General discussion	
<hr/>	
References	103
Summary	117
Acknowledgements	119
List of publications	122
About the author	123
Education statement	124



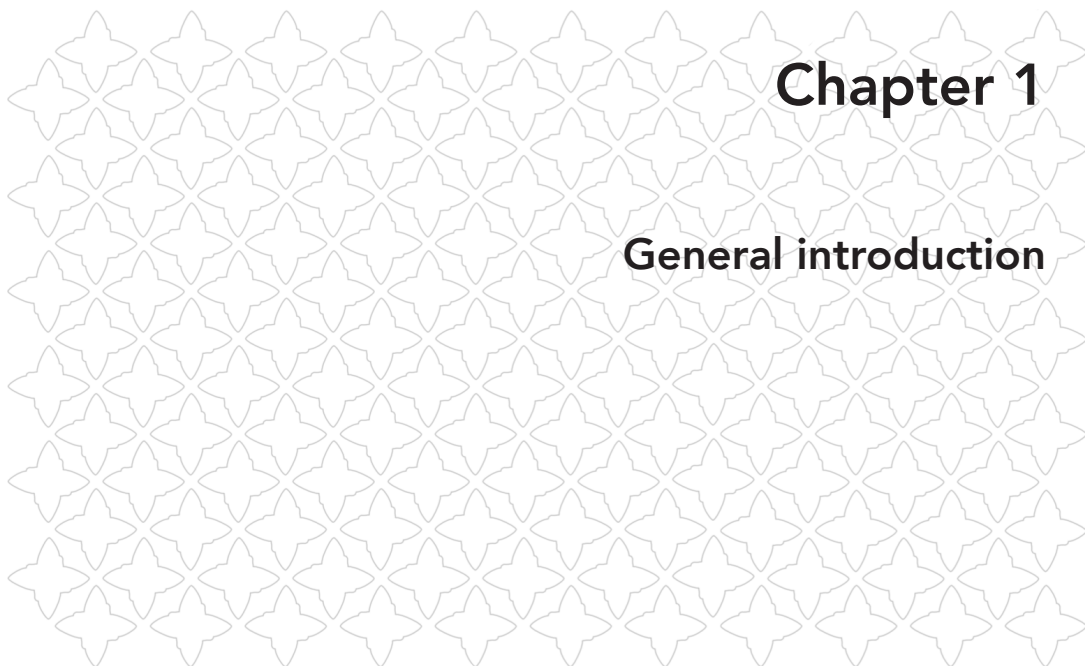


## Abbreviations

*A selection of abbreviations from this thesis.*

All-Lnc	Long non-coding RNA conserved at the nucleotide level by the Brassicaceae and Cleomaceae
Ath	<i>Arabidopsis thaliana</i> (rockcross)
Ath-Linc	Long intergenic non-coding RNA from Li <i>et al.</i> 2012
Ae-Lnc	Long non-coding RNA conserved at nucleotide level by 4 Aethionemeae species
Brass-Lnc	Long non-coding RNA conserved at the nucleotide level by <i>Arabidopsis thaliana</i> and <i>Aethionema arabicum</i>
CDS	coding sequence
Cleo-Lnc	Long non-coding RNA conserved at the sequence level by two Cleomaceae species
cM	centimorgan, unit used as the distance between two genes on a chromosome
GS	glucosinolates
LncRNA	Long non-coding RNA
LincRNA	Long intergenic non-coding RNA
LOD	logarithm of odds, used as a significance threshold in QTL analyses
MFE	Minimum Free Energy, used to assess the strength of a secondary structure of a RNA molecule
Mya	Million years ago
ORF	Open Reading Frame
QTL	Quantitative Trait Locus
RIL	Recombinant Inbred Line
SNP	Single Nucleotide Polymorphism
WGD	Whole Genome Duplication also known as polyploidisation





# **Chapter 1**

## **General introduction**

## Background

The economically important plant family Brassicaceae includes many crops (e.g. cabbage, canola, cauliflower and turnip) and the model plant *Arabidopsis thaliana*. Brassicaceae contains 3741 species in 325 genera (Al-Shehbaz 2012) and phylogenetic analyses have identified a larger core-group, with three main lineages, and a smaller sister-clade Aethionemeae (Beilstein *et al.* 2006, Beilstein *et al.* 2008, Couvreur *et al.* 2010, Franzke *et al.* 2009). More recently, the family has been split into lineages A-F (Huang *et al.* 2016). Polyploidy, or whole genome duplication (WGD), is a common feature of angiosperm evolution (Soltis *et al.* 2009, Soltis *et al.* 2010). The entire Brassicaceae shares a common whole genome duplication event referred to as At-alpha (Edger *et al.* 2015, Haudry *et al.* 2013). Brassicaceae are globally distributed over the temperate regions (Hohmann *et al.* 2015) and contain species that are metal-tolerant (e.g. *Noccaea caerulescens*), salt tolerant (e.g. *Eutrema halophila*), drought tolerant (e.g. *Allysum montanum*) and cold tolerant (e.g. *Draba chionophila*). Important synapomorphies of the Brassicaceae are not restricted to morphological characters like dissymmetric flowers, tetradynamous stamens, ovary with a false septum etc. but also include the ecologically important secondary defence compounds: the methionine-derived glucosinolates (Edger *et al.* 2015, Stevens 2001).

Being sister to the Brassicaceae core-group sets Aethionemeae at an evolutionary important position for comparative analysis of genome and trait evolution. The monotypic tribe Aethionemeae W. Aiton consists only of the genus *Aethionema*. *Aethionema* species occur mainly at the potential Brassicaceae centre of origin: the Irano-Turanian region (Al-Shehbaz 2012, Couvreur *et al.* 2010, Franzke *et al.* 2011, Warwick *et al.* 2010, Hedge 1976). The Irano-Turanian region harbours three major global biodiversity hotspots (Manafzadeh *et al.* 2016). *Aethionema* species occur on steep (mostly rocky) slopes between 500-3000m, have simple sessile leaves, are annual or perennial, can be diploid or polyploid and have mainly dehiscent fruits, although heterocarpism also occurs (Davis 1965, Hedge 1965), and one nickel hyperaccumulator, *Ae. spicatum* (Reeves & Adigüzel 2008). With only 61 species, *Aethionema* is much smaller than the Brassicaceae core-group. This asymmetrical radiation pattern, a small sister group to a large core group having a shared WGD event, occurs in many of the large plant families (e.g. Poaceae, Solanaceae, Schranz *et al.* 2012). Although hypothesized in the past (Soltis *et al.* 2009) it is now known that WGD did not cause this asymmetrical pattern (Haudry *et al.* 2013, Edger *et al.* 2015, Schranz *et al.* 2012, Tank *et al.* 2015). Schranz *et al.* (2012) hypothesized that the lag time between a WGD event and the radiation of the core group could explain the asymmetrical radiation pattern.

The important evolutionary position of *Aethionema* was recognized after a genetic study done on the chloroplast *rbcl* gene in 1994 (Price *et al.* 1994). Before the advances of molecular phylogenetics, analyses of morphological characters led to the placement of Aethionemeae species in the core group (Prantl 1891, Hayek 1911, Schulz 1936). However, the characters used, e.g. fruit morphology and trichome form, were not homologous (Beilstein *et al.* 2006, Beilstein *et al.* 2008). Hence, not only Aethionemeae but also other Brassicaceae clades have been revised and replaced using advances in molecular phylogenetics (Al-Shehbaz 2012). The tribe Aethionemeae used to consist of two genera, *Moriera* and *Aethionema*. After renaming the single *Moriera* species, *Moriera spinosa*, to *Aethionema spinosum* (Prantl 1891, Komarov 1934, Al-Shehbaz 2012) the tribe Aethionemeae became monotypic and consisting only of the genus *Aethionema*.

The annual *Aethionema arabicum* (L.) Andr. ex DC is an emerging model species for ecologically and evolutionarily interesting phenotypes, e.g. heterocarpism (Lenser *et al.* 2016). Heterocarpic plants have two sets of fruits on the same infructescence: dehiscent fruits with 1-4 seeds, that fall apart at the mother plant and indehiscent fruits, with a single seed, that are spread as such (Mühlhausen *et al.* 2013). Heterocarpism makes it possible to adopt different distribution strategies in variable environments (Lenser *et al.* 2016). *Aethionema arabicum* occurs, like the other *Aethionema* species, on rocky steep slopes in Cyprus, Iran and Turkey and has recently also been found in Bulgaria (Velchev 2015). Due to the very hot summers and cold winters in the Irano-Turanian region, *Ae. arabicum* has its complete life-cycle from the late spring towards early summer (Bibalani 2012). The release of the *Ae. arabicum* genome and transcriptome (Haudry *et al.* 2013, Edger *et al.* 2015) has enabled researchers to incorporate *Ae. arabicum* into a vast set of studies on genetic and genomic evolution. For example, it was found that the *Ae. arabicum* telomerase template domain has less than 25% sequence similarity to that of *A. thaliana* (Beilstein *et al.* 2012). These studies give us an insight in the genomic evolution of *Ae. arabicum* but do not include other aspects of its biology.

### **Problem description and research approach**

The road to understand how the whole genome duplication lag-time occurred is not straightforward. It combines all the variables that are at play for the evolution of a plant family, here namely the Brassicaceae. Any path to unravel such a large and complex question requires knowledge about the small-sister clade, *Aethionema*, from different perspectives and evolutionary levels. In this thesis, I attempt to do so by analysing diverse topics: from long non-coding RNAs, to phylogenetics, genome wide diversity analysis to the evolution of defence compounds.

Historically genomic non-coding regions have been seen as junk DNA and relics that natural selection was not able to clean up (Dinger *et al.* 2008, Palazzo & Gregory 2014 and the references therein). The last decades have shown that the genomic non-coding parts can regulate gene expression (Zhang *et al.* 2013). Identifying these elements makes it possible to have a broader look at the regulatory mechanisms of a genome. Putting these elements in an evolutionary framework enables the understanding of the level of conservation, or species specificity, of elements that eventually are able to influence plant phenotype (Wierzbicki 2012). Recently long non-coding RNAs have been identified in plants (Swiezewski *et al.* 2009). Long non-coding RNAs are non-translated, >200nt long mRNAs that can regulate gene-expression (Dinger *et al.* 2008). They can influence the flowering time of *Arabidopsis thaliana* (e.g. COOLAIR and COLDAIR, Ietswaart *et al.* 2012), are involved in seed dormancy in *A. thaliana* (Fedak *et al.* 2016) and can influence male sterility in long day rice accessions (Ding *et al.* 2012a). Although all these studies identified the conservation of long non-coding RNAs by sequence similarity they did not take the positional conservation of sequences into account.

The Irano-Turanian distribution of *Aethionema* species makes it a hard to collect field material due to the current political climate. Combining next generation sequencing techniques and natural history collections makes it possible to access much larger amounts of species and accessions. Natural history collections have been used to understand the ancestral areas of

human developed crops, the worldwide movement of these crops and their pests (Ames & Spooner 2008). Natural history collections can also contain species that have been encountered only once (Sebastian *et al.* 2010), are now extinct (Zedane *et al.* 2016), or occur in areas that are currently hard to reach due to political instability. While in the past mainly single markers were used to reconstruct phylogenetic relationships, next generation sequencing provides a wealth of data. Using next generation sequencing on herbarium specimens gives the opportunity to generate data from taxa that might otherwise be hard or impossible to reach. The presence of digitalised herbarium specimens in large databases, e.g. GBIF (2012), makes it possible to access specimens collected and kept all over the world. These databases make it feasible to answer questions on historical biogeography. Combining the evolutionary history of *Aethionema* with historical biogeography in a time-calibrated manner can tell us the geohistorical events that *Aethionema* has been through.

The field of population genetics and genomics compares different populations of a single species to assess: population structure, genetic diversity and whether or not such a structure depends on any underlying trait or depends on its environment (Luikart *et al.* 2003, Givnish 2010). This could also lead to environmental factors that might be important. The level of population structure can provide answers to the levels of migration and speciation that are now taking place. Transcriptome data can help us to understand the genome wide genetic diversity of *Ae. arabicum* accessions at the genomic functional level. As speciation and natural selection act on the level of phenotypes the difference between populations should also be addressed at the level of ecologically important phenotypes.

Natural selection and speciation act at the level of traits. The fitness of a plant and hence the probability of reaching the next generation not only depends on the fertility and flowering-time but also on how plants cope with herbivores. In a race against their herbivores, Brassicales developed glucosinolate defence metabolites (Edger *et al.* 2015). Glucosinolates (i.e. mustard oils) form a two-component plant defence with their associated myrosinase enzymes (Halkier & Gershenzon 2006). Glucosinolates and myrosinases are spatially separated in a plant cell and form upon contact with each other nitriles and isothiocyanates (Halkier & Gershenzon 2006, Koroleva *et al.* 2000). While all Brassicales contain glucosinolates the highest diversity of 120 different glucosinolates is found within the Brassicaceae (Edger *et al.* 2015, Halkier & Gershenzon 2006). The quality and the quantity of mustard oils vary through the plants life stage and depend on the accession (Kliebenstein *et al.* 2001). The release of the *Ae. arabicum* genome has already made it possible to use a bottom-up approach to assess the genes that are known from *A. thaliana* to play a part in the glucosinolate pathway (Hofberger *et al.* 2013). However, they do not tell what the quantity and quality of *Ae. arabicum* glucosinolates are, or which genomic locations are involved for glucosinolate regulation. A series of glucosinolate measurements through development of *Ae. arabicum* and a mapping population make it possible to investigate the polymorphisms and genomic locations of glucosinolate biosynthesis in *Ae. arabicum*.

Combining different approaches including phylogenomics and historical biogeography, the conservation of non-coding regions, population genomics, and the evolution of an ecologically important trait, is a major challenge. The combination of these different techniques makes it possible to look at the evolution of a species and clade from different angles and hopefully will lead to putting the 'eco' back into 'ecogenomics'.

## Research questions and objectives

This thesis approaches the genus *Aethionema* from different perspectives. By studying inter-species diversity, within species genetic diversity and the evolution of ecologically important traits, I present a more complete picture and understanding of *Aethionema* evolution. To accomplish that goal this thesis touches upon the following questions:

- a) Are there lineage-specific long non-coding RNAs for *Aethionema* and the Brassicaceae-core?
- b) What are the species relationships within *Aethionema*?
- c) What is ancestral area of *Aethionema*?
- d) What is the genetic diversity of *Aethionema arabicum*?
- e) What is pattern of glucosinolate variation along the development of *Aethionema arabicum*?

This thesis tries to answer these questions in a comprehensive way using current sequencing, bio-informatics and analysis approaches.

## Thesis outline

I have organized the six chapters of this thesis based on a focal continuum from macro-evolution, to micro-evolution and finally to trait evolution (Fig. 1), although every chapter can stand on its own. They are enclosed by this general introduction (Chapter 1) and a general discussion (Chapter 6) bringing everything together.

Chapter 2 presents the long non-coding RNAs that are conserved by position but not by sequence within the Brassicaceae and Cleomaceae. With a bio-informatical pipeline based on synteny I recovered the expressed long non-coding RNAs of *Aethionema*, core-group Brassicaceae and Cleomaceae. This chapter exemplifies the importance of the spatial conservation of genes in comparative genomic studies.

Chapter 3 addresses the phylogenetic relationships between the different *Aethionema* species and reconstructs the likely ancestral area in a time-calibrated framework. Using sequence data derived mainly from herbarium specimens, gave us the opportunity to cover 75% of all *Aethionema* species. With 76 chloroplast coding regions and the nuclear ribosomal regions the backbone of the phylogeny of *Aethionema* was resolved and the ancestral area of *Aethionema* assessed.

Wanting to understand what happens at a micro-evolutionary level, I used a population genomics approach in chapter 4 to assess the genetic diversity and signs of selection within the emerging model species *Ae. arabicum*. Using population genomic tools on transcriptomes of pooled tissue of seven individuals showed a clear population structure based on the underlying genomic variation. This population structure is also shown in the ecologically important traits of glucosinolates (and leaf shape).

Chapter 5 dives deeper into the ecologically important glucosinolate secondary metabolites. With a developmental-time series of two *Ae. arabicum* accessions, the change of glucosinolate quality and quantity through plant development is measured. Quantitative trait loci (QTL) mapping of a Recombinant Inbred Line population (RILs) derived from these two accessions

## Chapter 1

1

allowed us to identify the genomic locations controlling glucosinolate profiles of leaves, fruits and seeds. This allowed us to link glucosinolate profiles and life history transitions.

The general discussion, Chapter 6, provides a synthesis of all the above chapters and attempts to answer the questions postulated. The various topics are discussed in a broader range. These vary from the role of WGD on speciation, the importance of herbarium genomics for phylogenomics and population genomics and the understanding of trait evolution. Future perspectives and possibilities to follow up on these subjects are presented based on the findings of this thesis.

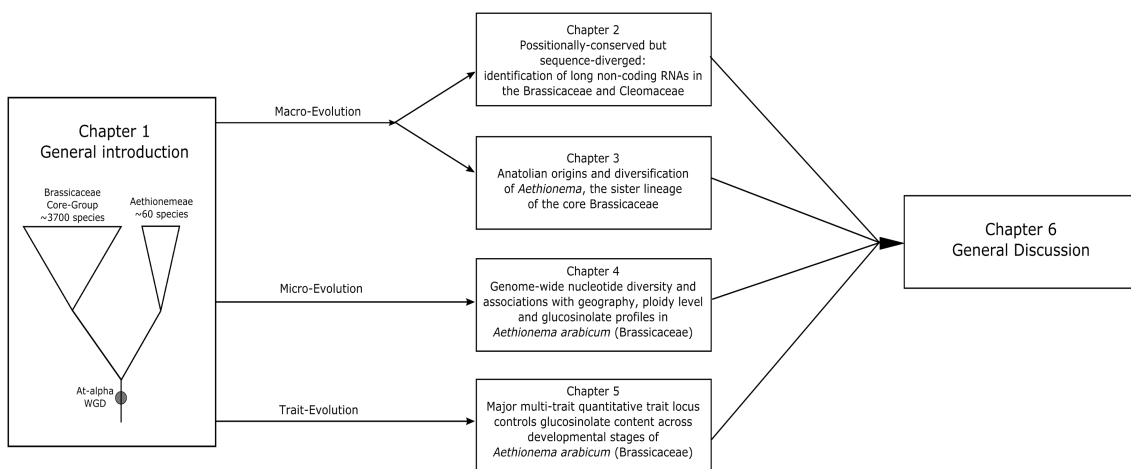


Fig. 1. Thesis structure





## Chapter 2

### Positionally-conserved but sequence-diverged: Identification of Long non-coding RNAs in the Brassicaceae and Cleomaceae

Setareh Mohammadin<sup>1</sup>, Patrick P. Edger<sup>2</sup>, J. Chris Pires<sup>3</sup>, M. E. Schranz<sup>1</sup>

Published in *BMC Plant Biology*, 2015, 15; 217-229.

DOI: 10.1186/s12870-015-0603-5

---

<sup>1</sup> Biosystematics, Plant Sciences Group, Wageningen University and Research, Wageningen, The Netherlands

<sup>2</sup> Department of Horticulture, Michigan State University, East Lansing, MI 48823, USA

<sup>3</sup> Division of Biological Sciences, University of Missouri, Columbia, Missouri 65211, USA

## Abstract

**Background** Long non-coding RNAs (LncRNAs) have been identified as gene regulatory elements that influence the transcription of their neighbouring protein-coding genes. The discovery of LncRNAs in animals has stimulated genome-wide scans for these elements across plant genomes. Recently, 6480 LincRNAs were putatively identified in *Arabidopsis thaliana* (Brassicaceae). However, there is limited information on their conservation.

2

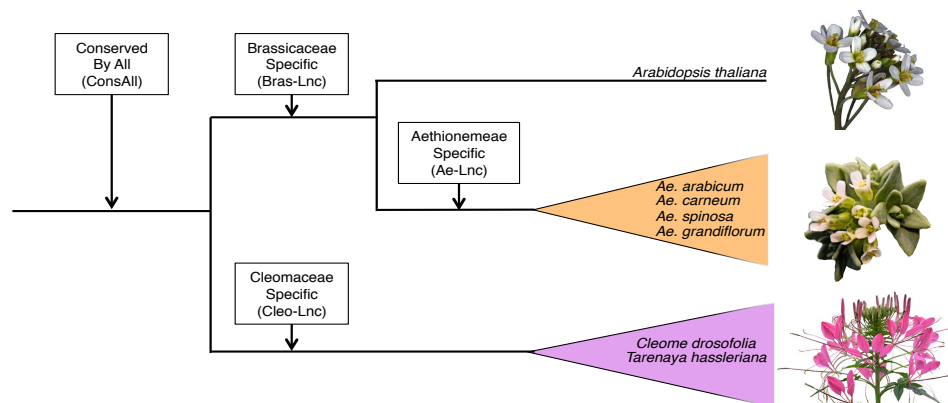
**Results** Using a phylogenomics approach, we assessed the positional and sequence conservation of these LncRNAs by analysing the genomes of the basal Brassicaceae species *Aethionema arabicum* and *Tarenaya hassleriana* of the sister-family Cleomaceae. Furthermore, we generated transcriptomes for another three *Aethionema* species and one other Cleomaceae species to validate their transcriptional activity. We show that a subset of LncRNAs are highly diverged at the nucleotide level, but conserved by position (syntenic). Positionally conserved LncRNAs that are expressed neighbour important developmental and physiological genes. Interestingly, >65 % of the positionally conserved LncRNAs are located within 2.5 mb of telomeres in *Arabidopsis thaliana* chromosomes.

**Conclusion** These results highlight the importance of analysing not only sequence conservation, but also positional conservation of non-coding genetic elements in plants including LncRNAs.

## Background

Gene regulatory transcripts are crucial in expressing or repressing protein coding genes. For example, gene repression in plants can be maintained by microRNAs (miRNAs, 19-22nt long) and small interfering RNAs (siRNAs, 23-24nt long). While miRNAs are mainly involved with the post-transcriptional gene repression, siRNAs are also involved pre-transcriptional gene repression by the *de novo* deposition of chromatin marks (Axtell 2013). A new category of RNA dependent gene regulators are Long non-coding RNAs (LncRNAs, longer than 200nt, ORF smaller than 100 amino acids) that can act in the course of pre-transcriptional repression of gene-expression (Zhang *et al.* 2013, Wierzbicki 2012, Dinger *et al.* 2008).

Long non-coding RNAs can silence genes by acting as a sequence-specific template for chromatin or associate with downstream proteins (Wierzbicki 2012) and are transcribed from the intergenic (long intergenic non-coding RNAs = LincRNAs), intronic or anti-sense regions (Liu *et al.* 2012, Zhang & Chen 2013). Recently it has been shown for the LncRNAs COOLAIR in *Arabidopsis thaliana* (Csorba *et al.* 2014, Swiezewski *et al.* 2009) and for the rice LncRNA LDMAR (Ding *et al.* 2012a, Ding *et al.* 2012b) how they influence the expression of phenotypically important regulatory genes. COOLAIR (cold induced long antisense intragenic RNA) is transcribed from the *Flowering Locus C (FLC)* and accelerates the transcriptional repression of *FLC* during cold by reducing the gene activating chromatin mark H3K36me3 (Csorba *et al.* 2014). In parallel, the gene silencing chromatin mark H3K27me3 is accumulating at the intragenic *FLC* nucleation site by a Polycomb-directed process (Csorba *et al.* 2014). Thus, LncRNAs COOLAIR contributes to the induction of flowering after vernalization. The mutant rice 58S has infertile pollen under long days, while the pollen are variably fertile under short days. Ding *et al.* (2012a) found that when LncRNA LDMAR is overexpressed in 58S rice recovers fertility under long days. The transcription of LDMAR in 58S is controlled by a negative feedback loop with a siRNA called Psi-LDMAR. Psi-LDMAR is transcribed from the promoter region of LDMAR. Psi-LDMAR induces RNA dependent DNA methylation; this leads to a reduction in the transcription of LDMAR and hence reduces the fertility of 58S under long days (Ding *et al.* 2012b). These recent discoveries of plant LncRNAs highlight their influence on important fitness traits, e.g. male sterility (LDMAR) and flowering time (COLDAIR, COOLAIR, IPS1) (Swiezewski *et al.* 2009, Ding *et al.* 2012a). The influence of



**Fig. 1** Simplified phylogeny of the Brassicaceae and Cleomaceae highlighting target species used to identify Long non-coding RNAs (LncRNAs). The boxes above the branches represent the studied lineages, their specificity at the sequence level and their abbreviations. Pictures show (from top to bottom) the inflorescences of *Arabidopsis thaliana*, *Aethionema arabicum* and *Tarenaya hassleriana*.

LncRNAs on regulating chromatin structure shows their involvement to permit plants to respond to environmental cues (Wierzbicki 2012).

LncRNAs have also been identified and studied in other plants, including *Zea mays*, *Triticum aestivum* and *Oryza sativa* (Li *et al.* 2007, Li *et al.* 2014, Xin *et al.* 2011). These genome-wide identifications of LncRNAs were done using existing EST sequences, full-length cDNA databases and/ or full genome tiling microarrays (Li *et al.* 2007, Li *et al.* 2014, Xin *et al.* 2011). Li *et al.* (2007) (Li *et al.* 2007) found more than 20,000 putative LncRNAs in rice; although >90% were assigned to being small RNA precursors. A similar result was found in *Zea mays* where ~60% of the LncRNAs are probably small RNAs precursors (Boerner & McGinnis 2012). About 40% of the rice non-exonic transcription active regions seem to be potential non-coding RNAs (Li *et al.* 2007). Liu *et al.* (2012) (Liu *et al.* 2012) found 6480 LincRNAs in the model plant *Arabidopsis thaliana* (Brassicaceae). Some of these putative L(i)ncRNAs were further validated with expression pattern analyses, custom microarrays and RNA-seq (Liu *et al.* 2012, Li *et al.* 2014, Xin *et al.* 2011, Li *et al.* 2007). However all these studies have thus far relied on analyses of only a single species.

Inter-species genome-wide comparisons have shown that protein-coding genes are not only conserved by sequence, but can also be conserved by their position in the genome (e.g. synteny) (Tang *et al.* 2008). The conservation of a genomic position over different phylogenetic scales can indicate that the position of a given gene is under strong purifying selection (Fridman & Zamir 2003). The genome-wide duplication history of *Arabidopsis thaliana* (Brassicaceae) was revealed by the identification and analyses of collinear duplicated blocks that arose from multiple ancient whole genome duplications (Bowers *et al.* 2003). Recently, the genome of *Aethionema arabicum*, a member of the Tribe Aethionemeae in the earliest diverging lineage of the Brassicaceae, was sequenced (Haudry *et al.* 2013) as well as the genome of *Tarenaya hassleriana* of the Cleomaceae, the sister-family to the Brassicaceae (Cheng *et al.* 2013). The comparisons of these three genomes provide insights into which genes and intergenic regions may be conserved by position between Brassicaceae-Cleomaceae. However, the genome sequences are not enough to understand their potential functional significance. Hence it is also valuable to have transcriptome data to complement the genome data of species at evolutionary important positions to infer the positional conservation of regulatory transcripts including LncRNAs.

Here we used the genomes of *Ae. arabicum*, *T. hassleriana* and *A. thaliana* in addition to our newly generated transcriptome data of four Aethionemeae and two Cleomaceae species to understand the conservation of LncRNAs in a phylogenomic context (Fig. 1). We not only analysed the nucleotide conservation of LncRNAs, but also whether or not they were conserved by genomic position. We found that of the LncRNAs that seem sequence-specific (e.g. lineage-specific) to the Cleomaceae, Brassicaceae or Aethionemeae, > 25% are conserved by position. This positional conservation could tell us more about the putative function of these LncRNAs, and the evolutionary importance of positional conservation of these genomic features.

## Materials and Methods

### Transcriptome isolation, library preparation and assembly

*Aethionema arabicum*, *Ae. carneum*, *Ae. spinosa*, *Ae. grandiflorum*, *Tarenaya hassleriana* and *Cleome droserifolia* seeds were germinated in sowing soil and grown in the greenhouse at the University of Amsterdam (18°C at night, 20°C day temperature, 12hours light, 12hours dark). Table 1 shows the tissues used for RNA isolation. To decrease RNA degradation the tissues were ground in liquid nitrogen and RNA was immediately isolated using PureLink™ RNA mini kit (Ambion, Life Technologies Corporation, Carlsbad, CA, USA), followed by a DNase treatment using the TURBO DNA-free™ kit (Ambion), according to the manufacturers protocols. The RNA quality and quantity was checked on a 1% agarose gel stained with ethidium-bromide in a 1xTBE buffer and on a NanoDrop 1000© spectrophotometer (Thermo Fisher Scientific, Wilmington, DE, USA). The samples were dried with GenTegra™ (GenVault, Carlsbad, CA, USA) for shipment to the Sequencing Core of the University of Missouri-Columbia. The ds-cDNA library was constructed following the manufacturers protocol of the TruSeq-RNATM kit (Illumina, San Diego, CA, USA). The six new transcriptomes used here were selected for mRNA during the cDNA synthesis. Thus all the non-polyadenylated LncRNAs were not sequenced. *Aethionema grandiflorum* and *A. spinosa* were paired-end sequenced with the Illumina Hiseq2000 sequencer on 1x100bp lanes, with 3 lines per lane. The *Ae. arabicum* transcriptome was *de novo* assembled using Trinity (Grabherr *et al.* 2011). The *Ae. carneum*, *Ae. grandiflorum* and *Ae. spinosa* transcriptomes were assembled against the *Ae. arabicum* contigs with NextGene V2.17® (SoftGenetics, State College, PA, USA) with matching requirements of  $\geq 40$  bp and  $\geq 90\%$  similarity and  $\leq 20\%$  present mutations. For each line a consensus sequence was constructed with the following parameter settings: 90% minimum of aligned reads for homozygosity, 25% as the cut-off for aligned read to be heterozygous and 85% as the percentage of reads that are aligned for a homozygote indel.

**Table 1** Species and tissues used for RNA isolation.

Species	Tissues
<i>Aethionema arabicum</i> & <i>A. carneum</i>	Fruits, flowers, buds, apical meristem, leaves and side buds from fully grown plants, leaves and apical meristem from juvenile plants, and the whole seedling including the roots
<i>Aethionema grandiflorum</i> & <i>A. spinosa</i>	Meristem, leaves and young stems
<i>Tarenaya hassleriana</i>	Buds, open flowers and the apical meristem
<i>Cleome droserifolia</i>	Young leaves, roots and flowers

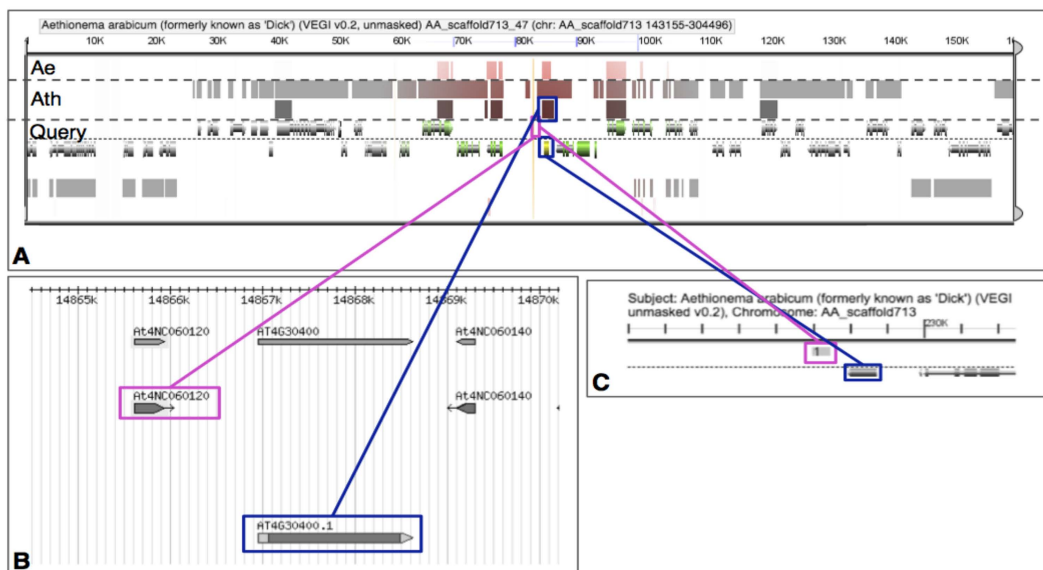
### Genomes, CDSs and LncRNA

The *Aethionema arabicum* and *Tarenaya hassleriana* genomes were downloaded from the CoGe Website (Cheng *et al.* 2013). The CDSs of *Brassica rapa*, *Arabidopsis lyrata* and *Eutrema halophila* come from the PlantGDB website (Duvick *et al.* 2008) and the *Arabidopsis thaliana* (Ath) CDS v10 from TAIR (Huala *et al.* 2001). The proteomes of *Zea mays*, *Oryza sativa*, *Brachypodium distachion*, *Sorghum bicolor* and *Sorghum italica* were downloaded from Phytozome (Goodstein *et al.* 2012). These latter CDS and proteomes were used in the OrthoMCL analysis (Suppl. Fig. 2) to ascertain that the LncRNAs are lineage specific. The location of Ath LncRNAs (Ath-Lnc) were downloaded from the PLncDB website (Jin *et*

*al.* 2013) and used to extract the sequences from the *A. thaliana* chromosomes (Huala *et al.* 2001) with an in-house python script. All the genomes present in November 2013 in Phytozome (Goodstein *et al.* 2012) were downloaded for latter analyses.

### OrthoMCL, BLAST and positional conservation analyses

OrthoMCL (Li *et al.* 2003), is based on reciprocal best BLAST hits (RBH) and uses a cluster algorithm (MCL) to cluster the RBHs. Depending on the blast that is performed it is possible to use OrthoMCL with nucleotide or protein sequences. We used OrthoMCL with blastN, query identity = 50% and  $evalue=1e-10$ , was used to assign orthologous groups to the lineage of interest (Suppl. Fig. 1 and 11). All BLASTs were done with command-line BLAST (Camacho *et al.* 2009) against the in-house made database of the Phytozome genomes and/or against the NCBI database with the ‘-remote’ command. The ORF size was assessed through the VirtualRibosome website (Wernersson 2006) for all six frames and with a strict start codon. The location of the *Ae. arabicum* transcripts and *T. hassleriana* transcripts to the nearest genes on their own genomes was assessed with CoGeBlast (Lyons & Freeling 2008). The *Ae. arabicum* unmasked genome v2.5 and *T. hassleriana* unmasked genome V4 were used. Only when the transcripts had a query hit of  $\geq 50\%$  and a HSP=1 they were assumed to hit to the correct location on the genome. Alternative splicing was excluded by this assumption, as is also the case for redundant genomic hits. SynFind and GeVo (Lyons & Freeling 2008) were



**Fig. 2** Example of collinearity and a positional conservation analysis of a Long non-coding RNA (LncRNA). A) Screenshot from GeVo. GeVo calculates the collinearity of a query sequence with the genome of a subject organism. The query here is the nearest protein-coding gene of *Ae. arabicum* shown in C, the subjects are *Ae. arabicum* and *A. thaliana*. Here there are two collinear regions in *A. thaliana*. The position of the positionally conserved LncRNA is shown with a pink box, while the protein coding genes of *A. thaliana* and *Ae. arabicum* are shown with blue boxes. B) Screenshot from the PLncDB website, shown are the *Arabidopsis thaliana* LncRNA (pink) and its nearest protein coding gene (blue). C) Screenshot from the CoGe Blast HSP. Pink is the *Aethionema arabicum* transcript along the *Ae. arabicum* genome. Blue is the nearest *Ae. arabicum* protein coding gene. This SynFind and GeVo analyses can be redone with the following link: <https://genomeevolution.org/r/fmnmf>

used to assess collinearity between the region of nearest protein coding gene of the LncRNA and the *A. thaliana*, *Ae. arabicum* and/or *T. hassleriana* genome(s). For example: if a protein coding gene of Ae-Lnc was collinear with a region in *A. thaliana* the 'GenomeBrowse' utility of PLncDB (Jin *et al.* 2013) was used to assess whether there was a LncRNA in the same direction (upstream, downstream or as a natural antisense) that corresponds with the location of the Ae-Lnc (Fig. 2). Hence these are LncRNAs different at the sequence level but are similar at position (see also Additional Fig. 11 for an counter example).

All transcripts were tested to see whether or not they could be micro-RNA precursors. To this end, they were blasted (BlastN) against the mirBase database (Griffiths-Jones *et al.* 2008). We used the RNAfold server (Gruber *et al.* 2008) to see whether the transcripts could have a stable secondary structure as a microRNA. The structure was assumed to be stable if the Gibbs free energy was between -30 and -80 kcal/mol.

### Conserved LncRNA and secondary structure

The conservation of LncRNA was tested according to the pipeline as described in Supporting Fig. 1. This was done with the 10%, 20% and 50% query identity for the OrthoMCL analyses at the beginning of the pipeline.

To assess whether the positionally conserved LncRNAs could have stable secondary structures the RNAalifold and RNAfold servers (Gruber *et al.* 2008) were used. RNAalifold uses aligned sequences of more than two species, while RNAfold calculates secondary structures based on a single RNA sequence. For the Ae-Lncs we used the transcripts of *Ae. arabicum*, *Ae. grandiflorum*, *Ae. carneum* and *Ae. spinosa*. Transcripts from the same species (if there present in the OrthoMCL analysis, see above) were used for the Brassicaceae specific LincRNAs. For the Cleomaceae specific LincRNAs of both *T. hassleriana* and *C. drosifolia* were used. To compare the positionally conserved LncRNAs the secondary structures of the Ath-Linc were also calculated.

## Results

### Sequence conservation

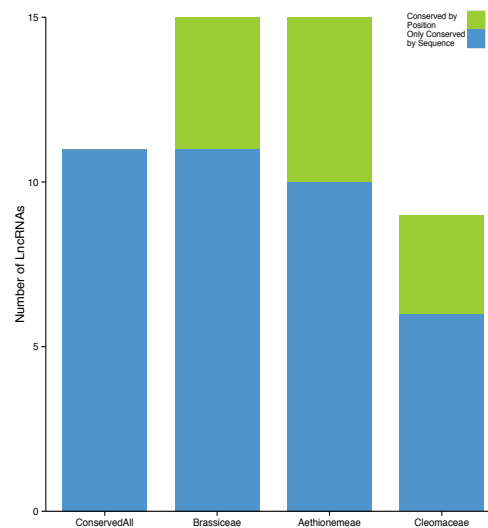
We identified LncRNAs in four Aethionemeae and two Cleomaceae species from transcriptome data. To assess the sequence conservation of these LncRNAs we used OrthoMCL (Li *et al.* 2003). For the positional conservation we used the CoGe tools SynFind and GeVo (Lyons & Freeling 2008).

We used a previous classification of LncRNAs in *Arabidopsis* (Liu *et al.* 2012): 1) LincRNA if the transcriptional unit (TU) was  $\geq 500$ bp away from the nearest protein coding gene, regardless if on sense or antisense strand. 2) Gene Associated Transcriptional Unit (GATU) if the TU was within a 500bp range of a protein-coding gene. 3) 'TU encoding NAT' if the TU was transcribed from the opposite strand than the sense strand of a protein coding gene. 4) miRNA precursors, which can have long transcripts as precursors.

We assessed whether the 6480 *A. thaliana* LincRNAs (Ath-Linc) assessed by (Liu *et al.* 2012) were conserved throughout the Brassicaceae and Cleomaceae (All-Linc) with an OrthoMCL analysis; a cluster algorithm based on reciprocal best blast hits (Li *et al.* 2003). The analysis included Ath-Linc and the genomes of *Aethionema arabicum* and *Tarenaya hassleriana* (see Materials and Methods and Suppl. Fig. 1 for details). Because LincRNAs have a higher mutation rate than protein coding sequences (Boerner & McGinnis 2012, Guttman *et al.* 2009), the analysis was done using increasing sequence similarity cut-off values of  $\geq 10\%$ ,  $\geq 20\%$  and  $\geq 50\%$ . Out of the 6480 Ath-Lincs only eleven are conserved by all three species at the genomic level. Out of these eleven conserved Ath-Lincs, only nine are transcribed in all three species based on our RNA-seq data (see below) and the RNA-seq data of (Liu *et al.* 2012) (Additional Table 1 for the average transcript and ORF lengths of these LincRNAs). Conserved Ath-Lincs were blasted (local BlastN) against the NCBI-database to assess whether the sequences were conserved in other organisms. At3NC056191, with a sequence similarity of  $\leq 20\%$  with the *Ae. arabicum* and *T. hassleriana* transcriptomes and genomes, was homologous in sequence to the 5.8S ribosomal RNA gene and internal transcribed spacer 2 to the oomycete *Albugo laibachii*. The genomically conserved At2NC003370, At4NC004390 and At4NC004390 were conserved across most land plants, including the bryophyte *Physcomitrella patens* (Suppl. Dataset).

We defined a lineage-specific LincRNA that is shared at the nucleotide level by multiple species within our focal lineages (e.g. Brassicaceae, Aethionemeae or Cleomaceae), but not found in other lineages. There were fifteen Ath-Lincs that were specific only to the Brassicaceae (Bras-Linc, Fig. 1). To ascertain that the Ath-Lincs and their corresponding *Ae. arabicum* transcripts were restricted to the Brassicaceae we compared them against the NCBI and Phytozome databases using BlastN, BlastX and TblastX (Material and Methods and Suppl. Fig. 1 for details and cut-off values). Of the fifteen Bras-Lincs, nine were transcribed by *Ae. arabicum* and/or *A. thaliana* (Suppl. Table 3 for the average transcript and ORF length of the *Ae. arabicum* transcripts).

To test for Aethionemeae specific LincRNAs (Ae-Linc) we generated RNA-seq data for four Aethionemeae species: *Ae. arabicum*, *Ae. carneum*, *Ae. grandiflorum* and *Ae. spinosa*. We identified 15 LincRNAs Ae-Lincs that were  $\geq 50\%$  similar in sequence between these four Aethionemeae species (Material and Methods and Suppl. Fig. 2 for pipeline). These fifteen Ae-Lincs correspond to 15, 15, 16 and 20 transcripts in *Ae. arabicum*, *Ae. carneum*,



**Fig. 3** Bar-plot of the number of lineage-specific Long non-coding RNAs (LncRNAs). Every bar shows the total number of LncRNAs that are conserved by sequence within that clade. The green bars are the number of LncRNAs that are conserved by position across every clade and the blue bars are conserved by sequence within their lineage. For example: out of the nine LncRNAs that are by sequence conserved within the Cleomaceae, three are conserved by position in *Arabidopsis thaliana* and six are lineage specific by sequence and position to the Cleomaceae.



*Ae. grandiflorum* and *Ae. spinosa* respectively (from the total of 19037, 18305, 48609 and 60772 predicted transcripts). The average ORF length ( $\pm$ SD) of the putative LncRNAs across all four species was 145.89 bp ( $\pm$  10.00 bp) with an average transcript length of 546.83 bp ( $\pm$  28.63 bp) (Suppl. Table 4 for species specific averages). The Ae-Lnc consisted of two GATUs, four TUs encoding NATs and nine LincRNAs (Suppl. Data and Suppl. Table 2). Two Ae-LncRNAs are micro-RNA precursors for ath-MIR403 and aly-MIR408 (MFE of -71.8 and -74.2 kcal/mol respectively). Although ath-MIR403 is not tissue specifically expressed, under hypoxic conditions it is more present in leaves and whole plants than in roots (Sunkar & Zhu 2004, Moldovan *et al.* 2010). The function and tissue specificity of aly-MIR408 is not known (Ma *et al.* 2010).

For the Cleomaceae-specific LncRNA (Cleo-Lnc), RNA-seq data of *Tarenaya hassleriana* and *Cleome droserifolia* were identically analysed as discussed above for the Ae-Lnc (Suppl. Fig. 2). We identified nine Cleomaceae-specific LncRNA based on 84,967 transcripts for *T. hassleriana* and 54,332 transcripts of *C. droserifolia* with  $\geq$ 50% sequence similarity. These nine transcripts had an average ORF and transcript-length ( $\pm$ SD) of 181.5 bp ( $\pm$ 7.78bp) and 675.71bp ( $\pm$ 201.53bp) respectively (Suppl. Table 3 for species specific lengths). According to the categorization mentioned above, these nine LncRNAs consist of two GATUs, four TUs encoding NATs and 3 putative LincRNAs. We did not identify any putative microRNA precursors.

### Conservation by position of transcribed LncRNAs

To exclude conserved non-coding sequences (CNSs) and to support functionality we only considered LncRNAs that we detected as transcribed by at least one species.

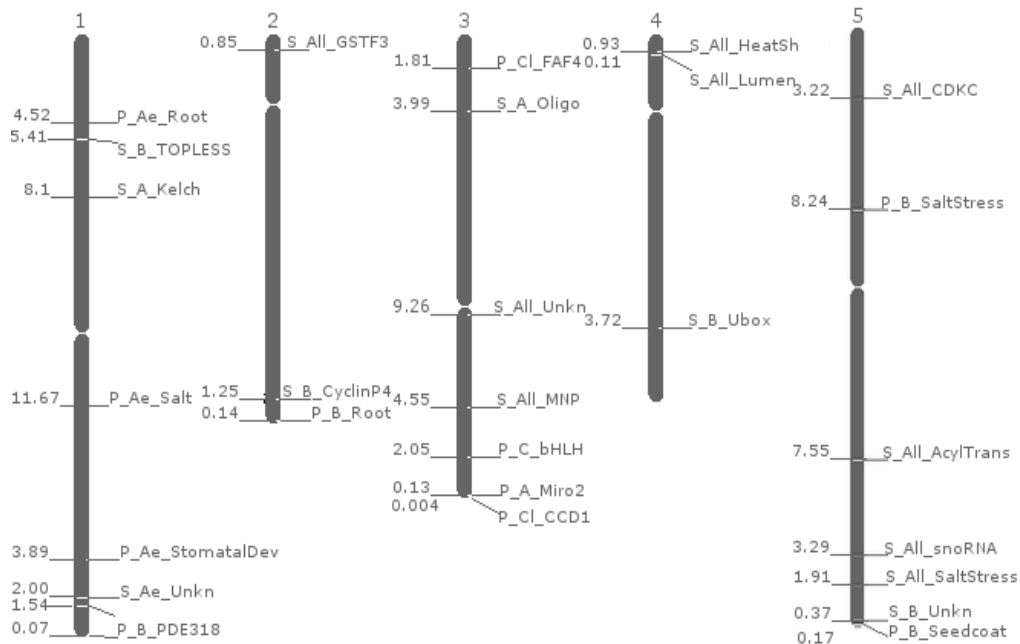
We analysed the transcribed lineage-specific LncRNAs per clade and whether or not they are conserved by position within the genome of another lineages. Positional conservation was assessed with the CoGe-tools CoGeBlast, SynFind and GeVo (Lyons & Freeling 2008), see Material and Methods for details). Out of the 39 LncRNAs that seemed to be lineage-specific at the nucleotide level (e.g. highly diverged between clades; 15 Bras-Lncs, 15 Ae-Lncs and 9 Cleo-Lncs) twelve were conserved by position in at least one of the other lineages (Fig. 2 for an example and Supporting Figures 3-9 for the others). Depending on the clade (Aethionemeae specific, Cleomaceae specific or Brassicaceae specific) the percentage of LncRNAs that are not conserved by sequence but are conserved by position in another clade varied between 26%-33% (Fig. 3 and Suppl. Table 2). Figure 4 shows the distribution of the positionally conserved LncRNAs as positioned in the *A. thaliana* genome. Remarkably 66.66% (8 out of 12) of the positionally conserved LncRNAs are within 2.5mb from the chromosome ends, including in the subtelomeric regions (Fig. 4 and Suppl. Table 2). This corresponds with the finding of others that the telomeres and subtelomeric regions, have a higher gene density than the genomic average (Bass & Birchler 2012). This could accordingly indicate the high number of gene regulatory elements.

Table 2 shows the functions of the neighbouring genes to the positionally conserved LncRNAs. The neighbouring genes of Bras-Lnc and Ae-Lnc (AT5G62420, AT5G24270 and AT1G50640) are associated with response(s) to salt stress. The *A. thaliana* genes neighbouring the positionally conserved Brass-Lnc and Ae-Lnc are involved at different

**Table 2** Function of the nearest protein coding gene in *Arabidopsis thaliana* of positionally conserved LncRNAs. 'Level of Sequence Conservation' denotes the level of lineage specificity of the LncRNA at the nucleotide level. Ae. Specific = Aethionemeae Specific; Brass. Specific = Brassicaceae (including Aethionemeae specific); Cleo Specific = Cleomaceae specific. Ae = Aethionemeae, Ath = *Arabidopsis thaliana*, Cleo = Cleomaceae.

Level of Sequence Conservation	Transcribed in	Abbreviation	<i>A. thaliana</i> LncRNA	<i>A. thaliana</i> Gene	Function
Brass. Specific	Ath	Bras_PDE318	At1NC112890	AT1G80770	Pigment defective 318 (PDE318)
Brass. Specific	Ath	Bras_Root	At2NC078030	AT2G47750	Morphological Effect: Root Growth Encodes GH3.9, a member of the GH3 family auxin-responsive genes. gh3.9-1 mutants had greater primary root length.
Brass. Specific	Ath	Bras_SaltStress	At5NC030470	AT5G24270	Response To Salt Stress. Encodes a calcium sensor that is essential for K+ nutrition, K+/Na+ selectivity, and salt tolerance
Brass. Specific	Ath	Bras_SeedCoat	At5NC103231	AT5G67180	Morphological Effect: Seed Coat Mucilage. Target of early activation tagged (EAT) 3 (TOE3)
Ae. Specific	Ae	Ae_NucIStruc	At1NC016180	AT1G13230	Required for growth promotion and enhanced seed production mediated by the endophytic fungus <i>Piriformospora indica</i> in <i>Arabidopsis</i> .
Ae. Specific	Ae	Ae_SaltStress	At1NC070280	AT1G50640	Response To Salt Stress and Involved in Leaf Senescence.
Ae. Specific	Ae	Ae_StomatalDev	At1NC099220	AT1G70410	Ethylene Responsive Element Binding Factor 3 (ERF3)
Ae. Specific	Ae	Ae_RepairPSII	Group1797	AT1G75690	Morphological Effect: Stomatal Development. Beta Carbonic Anhydrase 4 (BCA4)
Ae. Specific	Ae	Ae_Miro2	Group4790	AT3G63150	Physiological Effect: Repair of Photosystem II. Low Quantum Yield of Photosystem II 1 (LQY1)
Cleo. Specific	Cleo	Cleo_bHLH	Group4645	AT3G57800	Physiological Effect: Embryo Genesis and Mitochondrial Morphogenesis. Miro-Related GTP-ASE 2 (MIRO2)
Cleo. Specific	Cleo	Cleo_Unknown	Seed_Group2679	AT3G06020	Basic helix-loop-helix (bHLH) DNA-binding superfamily protein. Transcription Factor Family
Cleo. Specific	Cleo	Cleo_CCD1	Group4801	AT3G63520	Regulation of shoot meristem size (FAF4). Specifically expressed in vascular tissue. Carotenoid Cleavage Dioxygenase 1 (CCD1)

## Long non-coding RNAs in the Brassicaceae and Cleomaceae



**Fig. 4** Distribution of the Long non-coding RNAs (LncRNAs) across the *Arabidopsis thaliana* genome. The positions are named as follow: conservation\_level\_lineage of sequence conservation\_gene function. Conservation level can be P: conserved by position across multiple lineages. S: only conserved by sequence and not by position. Ae: conserved by sequence only in Aethionemeae. All: conserved by sequence through Brassicaceae and Aethionemeae. B: conserved by sequence only in Brassicaceae, including Aethionemeae. Cl: conserved by sequence only in Cleomaceae. The numbers left of the chromosome are the distances from the gene to the end of the chromosome in Mega bases.

levels of morphological and physiological development. These range from influencing root growth, to the development of stomata, to repairing photosystem II, to embryogenesis and mitochondrial morphogenesis (Table 2).

Some LncRNAs have been shown to have a stem-loop secondary structure (Ding *et al.* 2012a, Flintoft 2013, Novikova *et al.* 2012). We looked whether our positionally conserved LncRNAs have putative stable secondary structures and whether or not there are common features between the positionally conserved LncRNA (Fig. 5 and Suppl. Fig. 10). The stability of a secondary structure is determined by its Minimum Free Energy (MFE), assuming that the lower the energy, the more stable the structure is (Zuker & Stiegler 1981). Hence we regard structures with a MFE  $\geq -80$  kcal/mol as unstable. The secondary structures of the Ae-Lnc and their Ath-Linc counterparts are hence unstable (Fig. 5). The two Cleo-Linc and the Bras-Linc are more stable (Fig. 5). In accordance to the secondary structures found with other LncRNAs (Ding *et al.* 2012a, Flintoft 2013, Novikova *et al.* 2012) all the stable structures have long stems and big loops on one side (Fig. 5).

### Discussion

As more complete genomes become available, it is possible to use genetic collinearity in addition to sequence similarity to address questions of conservation of non-coding sequences in a phylogenomic context. Using a comparative approach with the sister families Brassicaceae and Cleomaceae, we found LncRNAs are positionally conserved and expressed,

but highly diverged at the nucleotide level. Hence here we found plant LncRNAs that are conserved by position but not by sequence, while the LncRNAs that are conserved by sequence are not conserved by position. While this result has been described earlier in comparative animal studies (Batista & Chang 2013), to the best of our knowledge our work represents the first example of this trend in plants.

Long (intergenic) non-coding RNAs have been shown to affect the expression of their neighbouring genes (Batista & Chang 2013), thus suggesting the importance of positional conservation in properly regulating adjacent genes encoding various traits. For example the positionally conserved LncRNAs found here are adjacent to genes involved in: response to salt stress, affecting important physiological functions (e.g. Photosystem II repair mechanism) or influencing morphological structures (e.g. root growth).

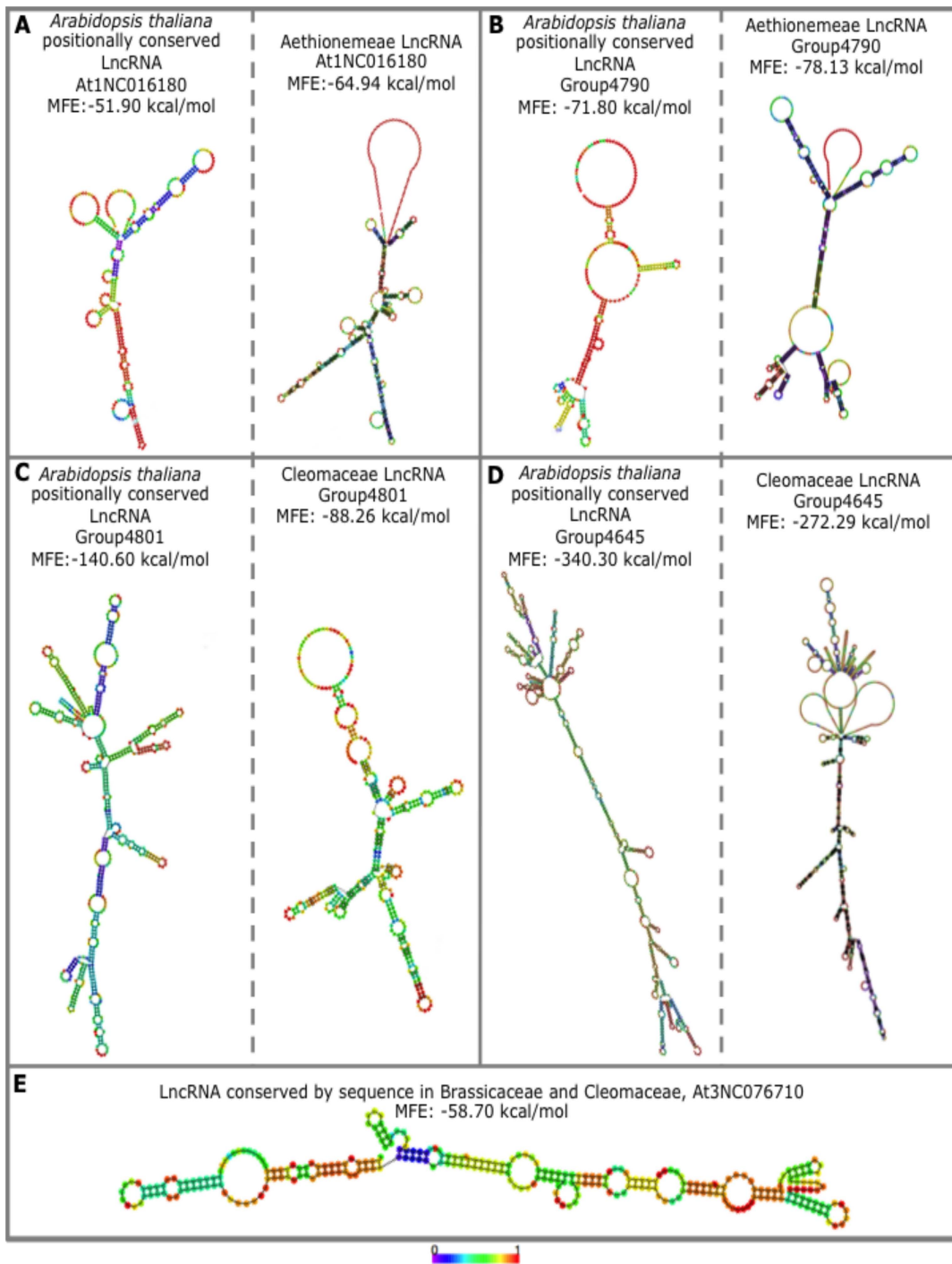
2

We based our analysis of positional conservation on the latest available genomes of *Aethionema arabicum*, *Tarenaya hassleriana* and *Arabidopsis thaliana*. The latest published *Aethionema arabicum* genome is >85% of its total genome size (Haudry *et al.* 2013) and the latest published genome of *Tarenaya hassleriana* is >94% of its total genome size (Cheng *et al.* 2013). Although these genomes have already been published our analyses are always limited by quality of the genome assembly.

Long non-coding RNAs are a potentially important feature of gene regulation and genomes of eukaryote organisms. To date, research into LncRNAs is more extensive in vertebrates than plants. Twenty-five out of the forty-eight functionally verified vertebrate LncRNAs have been conserved between human and mouse at >50% sequence similarity (Pang *et al.* 2006). Liu *et al.* (2012) (Liu *et al.* 2012), whose data has been explored here, found that <2% of all the putative LncRNAs they found in *A. thaliana* are conserved across the plant kingdom. A similar number has been found by comparing maize (monocot) LncRNAs and *A. thaliana* (eudicot) (Li *et al.* 2014). The LncRNAs of legumes show only 5% sequence conservation in non-legume plants (Wen *et al.* 2007). A much higher percentage of the *Zea mays* LncRNAs, <25%, are conserved in the closely related species sorghum (Li *et al.* 2014). Here we found that out of a total of 39 transcribed LncRNAs are diverged at the nucleotide level, twelve are conserved by position. This is more than 30% of the LncRNAs that we found in the transcriptomes of Aethionemeae and Cleomaceae.

Studies that take the position of LncRNAs into account primarily assume sequence conservation and additionally analyse whether or not those LncRNAs are also conserved by position. However in a comparison between zebra fish and humans Batista and Chang (2013) found that LncRNAs with weak sequence conservation can still be fully functional, because they are still structurally and positionally conserved. Here we show similar results in plants: positional conservation of LncRNAs with weak sequence similarity between distantly related species.

The lack of sequence conservation but the presence of positional conservation might be explained by an increase in mutation rate for these regulatory elements. This has already been pointed out by Pang *et al.* (Pang *et al.* 2006), who hypothesized, for miRNAs and longer non-coding RNAs, that the type of interaction within a regulatory network can be under selection pressure rather than the sequence of the regulatory element itself. This hypothesis



**Fig. 5** Secondary structures and Minimum Free Energy (MFE) of sequence and/or positionally conserved Long non-coding RNAs (LncRNAs). (A) LncRNAs that have both sequence conservation and positional conservation between *Arabidopsis* (left) and *Aethionema* (right). (B) LncRNAs that have only positional conservation between *Arabidopsis* (left) and *Aethionema* (right) (C) LncRNAs that have both sequence conservation and positional conservation between *Arabidopsis* (left) and *Tarenaya* (right) (D) LncRNAs that have only positional conservation between *Arabidopsis* (left) and *Tarenaya* (right) (E) The LncRNA conserved by sequence and position in *A. thaliana*, *Aethionema arabicum* and Cleomaceae. The colored bar below shows the basepairing probability for every structure.

would fit well with the regulatory function and the position of LncRNAs. As LncRNAs regulate the expression of their neighbouring protein coding gene their interaction with this gene, and hence their position, rather than their sequence can be under selection.

We compared the secondary structure of the positionally conserved LncRNAs (Fig. 5). In addition to the positional conservation of LncRNAs their secondary structure might also be conserved. The Aethionemeae positionally conserved LncRNAs are less stable (higher MFE) than the Cleomaceae positionally conserved. A similar difference in stability is seen in their positionally conserved counterparts in *Arabidopsis thaliana*. The stability of the LncRNA secondary structure might be a step to subdivide the big group of LncRNAs.

2

Genomic regions of different species can be similar in sequence and can be completely collinear. However, these sequences should not necessarily be transcribed (Suppl. Table 2). Here we used polyadenylated mRNAs to try to assess conservation of LncRNAs between different species. It has been shown that although LncRNAs can be polyadenylated, they are not always polyadenylated (Di *et al.* 2014). Consequently the positional conservation shows only a subset of the plants possible transcripts. Moreover we applied the stringent rule that every LncRNA had to be transcribed in at least two species from the same lineage. Hence these results in a set of highly confident positional conserved LncRNAs that represent only the tip of the iceberg.

The small number of conserved LncRNAs found here is in accordance with the findings in other systems as discussed above (Liu *et al.* 2012, Wen *et al.* 2007, Li *et al.* 2014). The consistent finding of low nucleotide conservation raises new questions about the mutation rate of LncRNAs. Studies have shown that the mutation rate of LncRNAs resembles those of introns (Mattick & Gagen 2001, Pang *et al.* 2006, Li *et al.* 2014, Batista & Chang 2013), which could partially explain the lack of sequence similarity between LncRNAs over deep evolutionary time. However, this lack of sequence similarity did not result in a lack of conservation by position, which could indicate a conservation of function as it has been shown earlier that positional conservation also accounts for functional conservation (Batista & Chang 2013, Tang *et al.* 2008).

The presence of more than 65% of the positionally conserved LncRNAs only within the 2.5 mb of chromosome arms is remarkable and unexpected. In many plants the sub-telomeric regions consist of repeats, called satellites though these are absent in *A. thaliana* (Bass & Birchler 2012). Their presence varies between species and even individuals within a species (Bass & Birchler 2012). The satellites in the sub-telomeric region typically consist of large A-T rich repeat stretches, which makes bending the DNA easier and the heterochromatin formation tighter, which is shown by the presence of dense heterochromatin blocks (Macas *et al.* 2002, Bass & Birchler 2012). One suggested function of the presence of these satellite arrays is their support of the chromatin states in the sub-telomeric region (Bass & Birchler 2012). However the absence of satellite arrays in *A. thaliana* might be compensated by the presence of LncRNAs that regulate the chromatin signatures of the protein coding genes in the sub-telomeric regions. We do not know of a specific reason why positionally conserved LncRNAs should be found only at chromosome ends. Certainly more research is needed to address this finding and the hypothesis stated above.

Preferably we would have tested whether the positionally conserved LncRNAs are also within 2.5 mb of the chromosome arms of *Aethionema arabicum* and/or *Tarenaya hassleriana*. However chromosomal-level genome assemblies of these species are not available yet. However, we are working on these genome assemblies so that we can address these questions in the near future.

Long (intergenic) non-coding RNAs have been identified by investigating the deleterious effects of knocking out these conserved sequences on various traits, e.g. flowering time, fertility, etc. (Swiezewski *et al.* 2009, Ding *et al.* 2012a, Zhang & Chen 2013). These wet-lab experiments are crucial to understand the functionality of any putative pathway (from gene and transcription to fitness effects). They can confirm the lack of small ORFs in LncRNAs and understand the full pathway on which the LncRNA has an effect, whether that is on neighbouring genes or across chromosomes (Hanada *et al.* 2007, Batista & Chang 2013).

## Conclusion

To summarize, we have shown here, using the Brassicaceae and Cleomaceae phylogenomic system that transcribed plant Long non-coding RNAs (LncRNAs) that seem to be only conserved within one lineage at the sequence level are conserved in other lineages at the same genomic position. The positional conservation could also imply a conservation of function but a divergence of sequence. Moreover, >65% of the positionally conserved LncRNAs are located within 2.5 mb of the telomeric region. This emphasizes the gene regulatory role that LncRNAs can have. These results imply that lineage specificity should not only be regarded at the nucleotide level but also at the positional level.

## Acknowledgements

This research was funded by the NWO Vernieuwings Impuls VIDI (Grant number: 864.10.001).

## Supplemental files

All Supplemental files are available upon request.

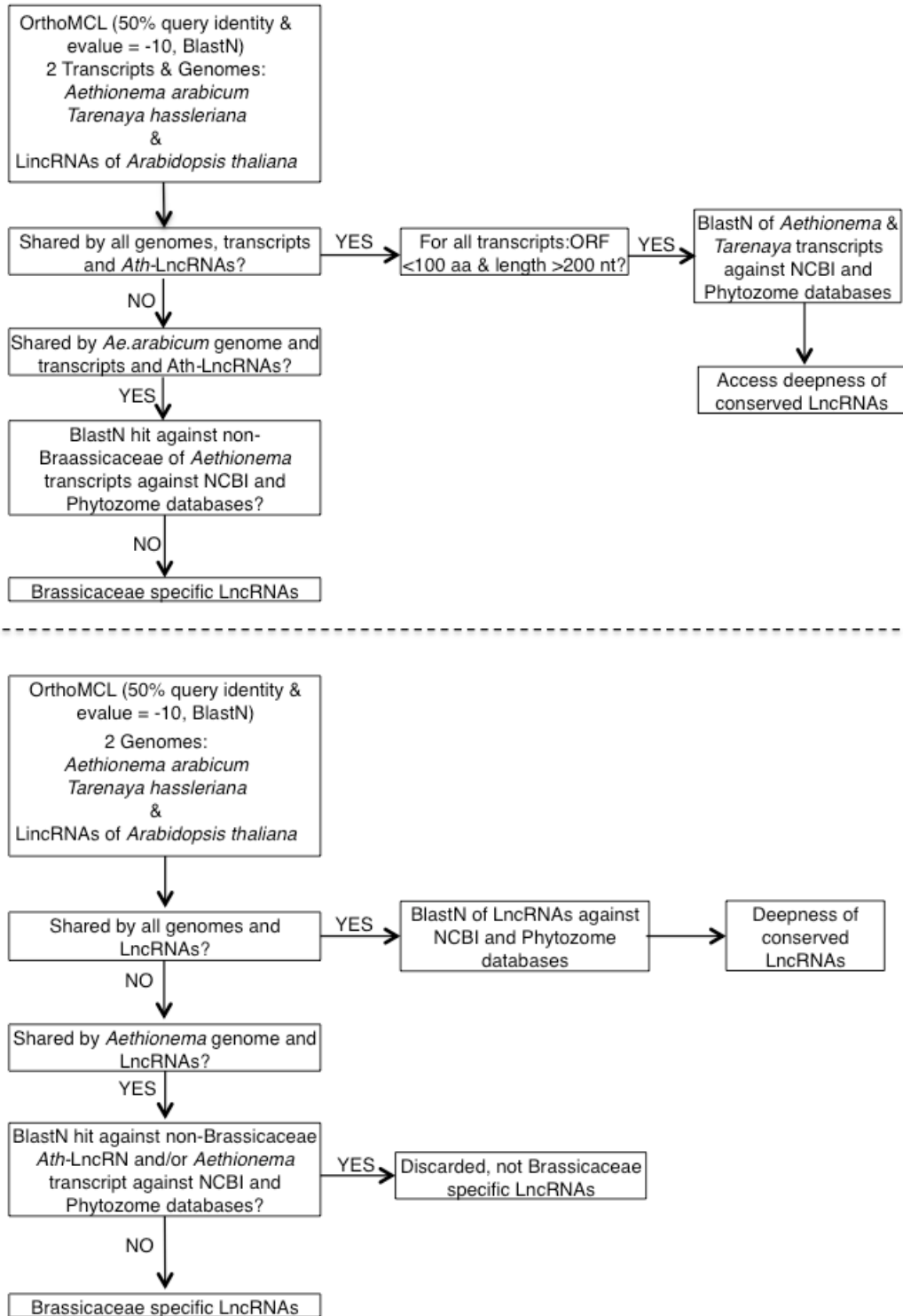
## Tables

Suppl. Table 1. Transcript and ORF length of *Tarenaya hassleriana* and Aethionemeae transcripts conserve by sequence. The sequence similarities percentages are cut-offs of sequence similarity within OrthoMCL.

Suppl. Table 2. Sheet 1 shows the transcript names of *Aethionema arabicum*, *Tarenaya hassleriana* and *Arabidopsis thaliana* and whether or not they are conserved by position or only by sequence across the different lineages. Sheet 2 shows the distances of the positionally conserved LncRNAs to the nearest end of *A. thaliana* chromosomes.

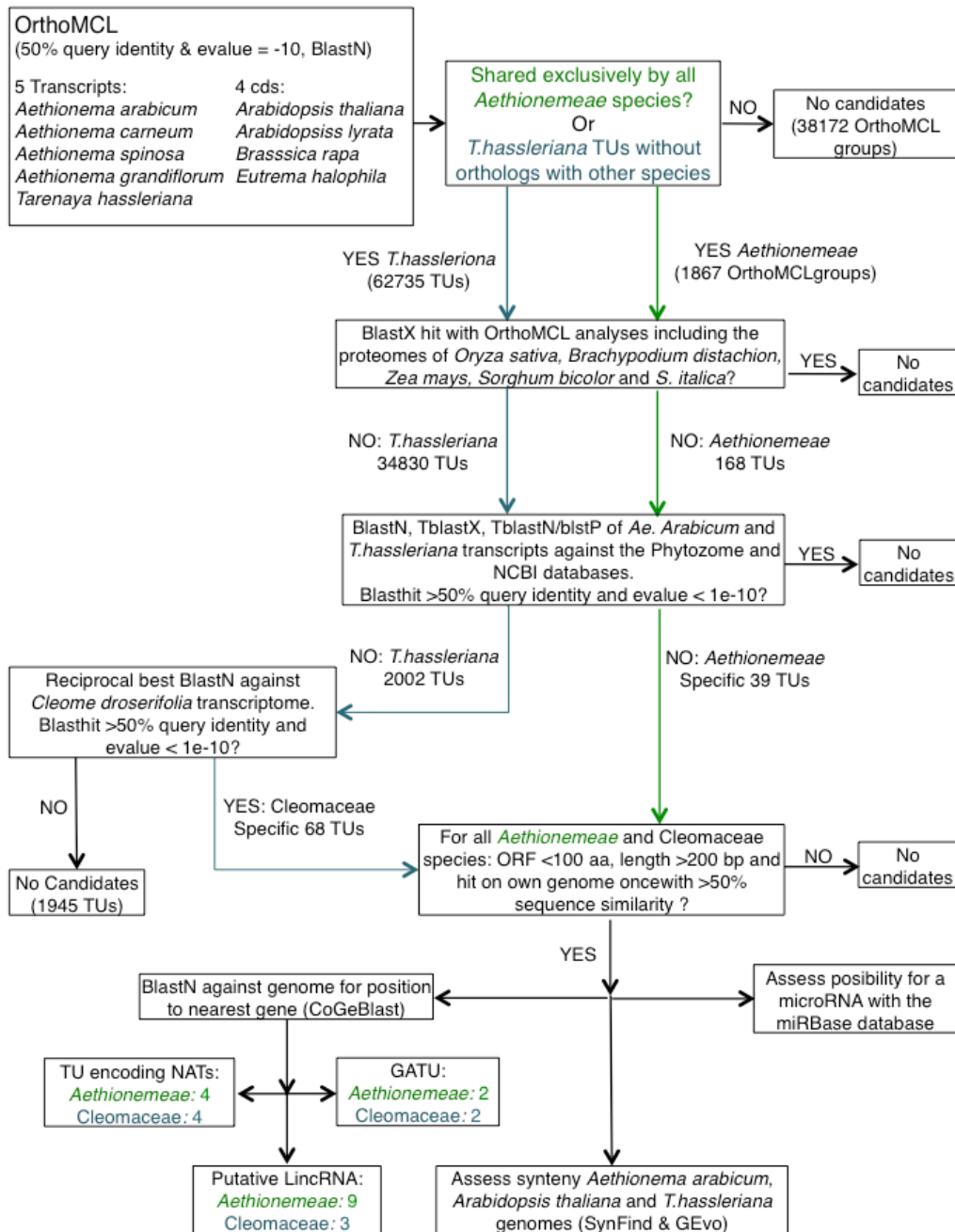
Suppl. Table 3. Transcript and ORF length of Aethionemeae and Cleomaceae specific Long non-coding RNAs.

Suppl. Table 4. Transcript and ORF length of Aethionemeae transcripts that are Brassicaceae specific. The sequence similarities percentages are cut-offs of sequence similarity within OrthoMCL.



**Suppl. Fig. 1** Pipeline to assess the transcribed (top-panel) and genomic (bottom-panel) Long non-coding RNAs (LncRNA) that are conserved at the nucleotide level throughout the Brassicaceae and Cleomaceae, or are specific to the Brassicaceae.

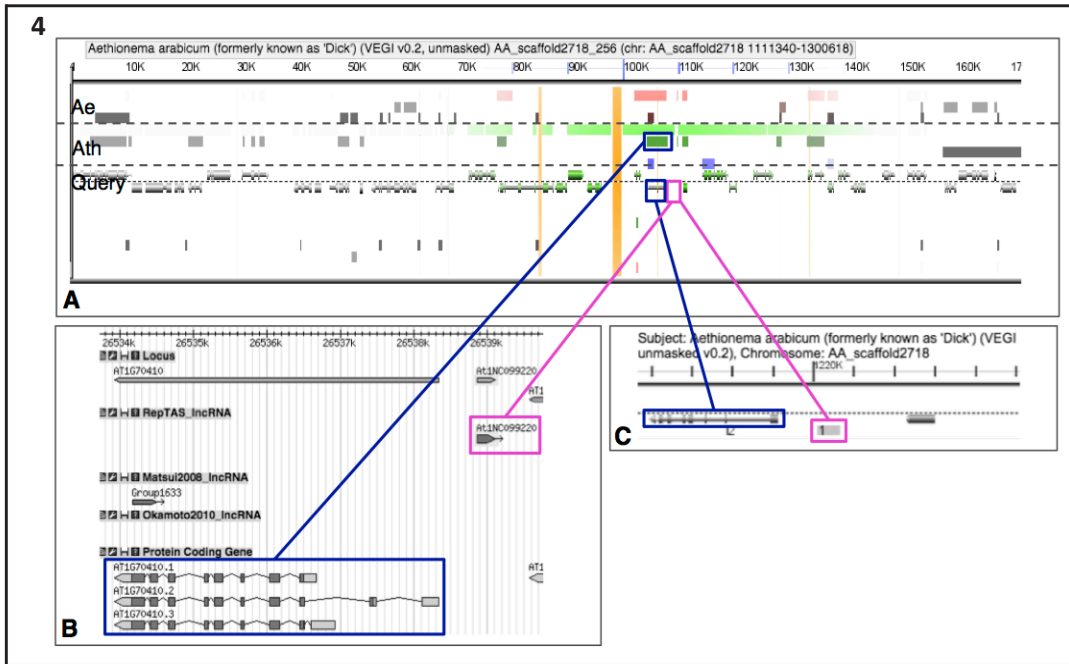
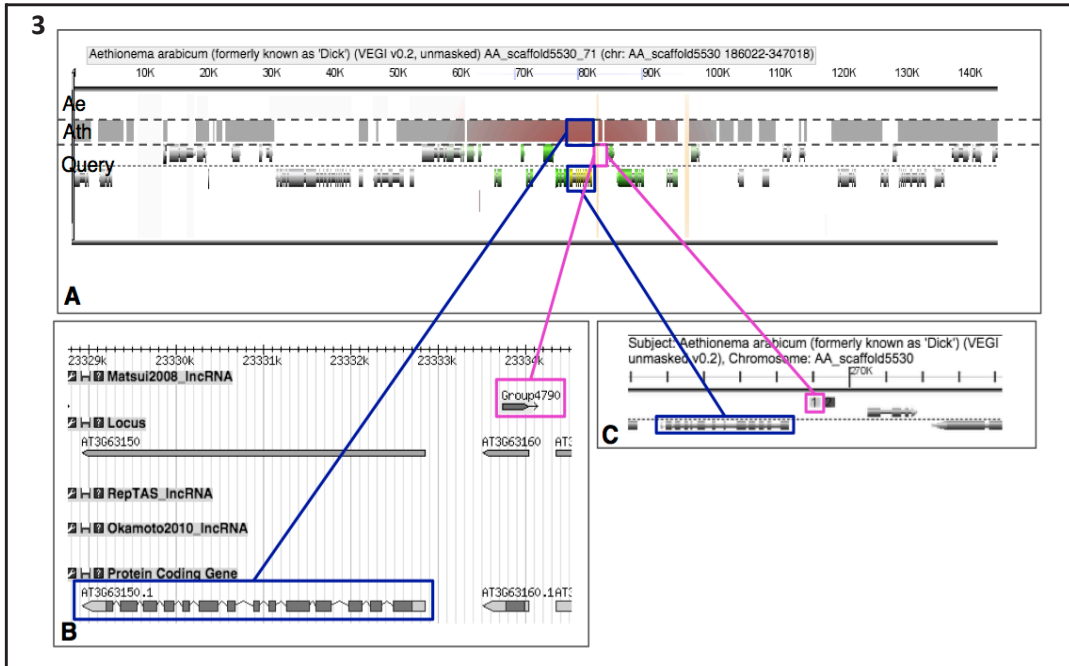


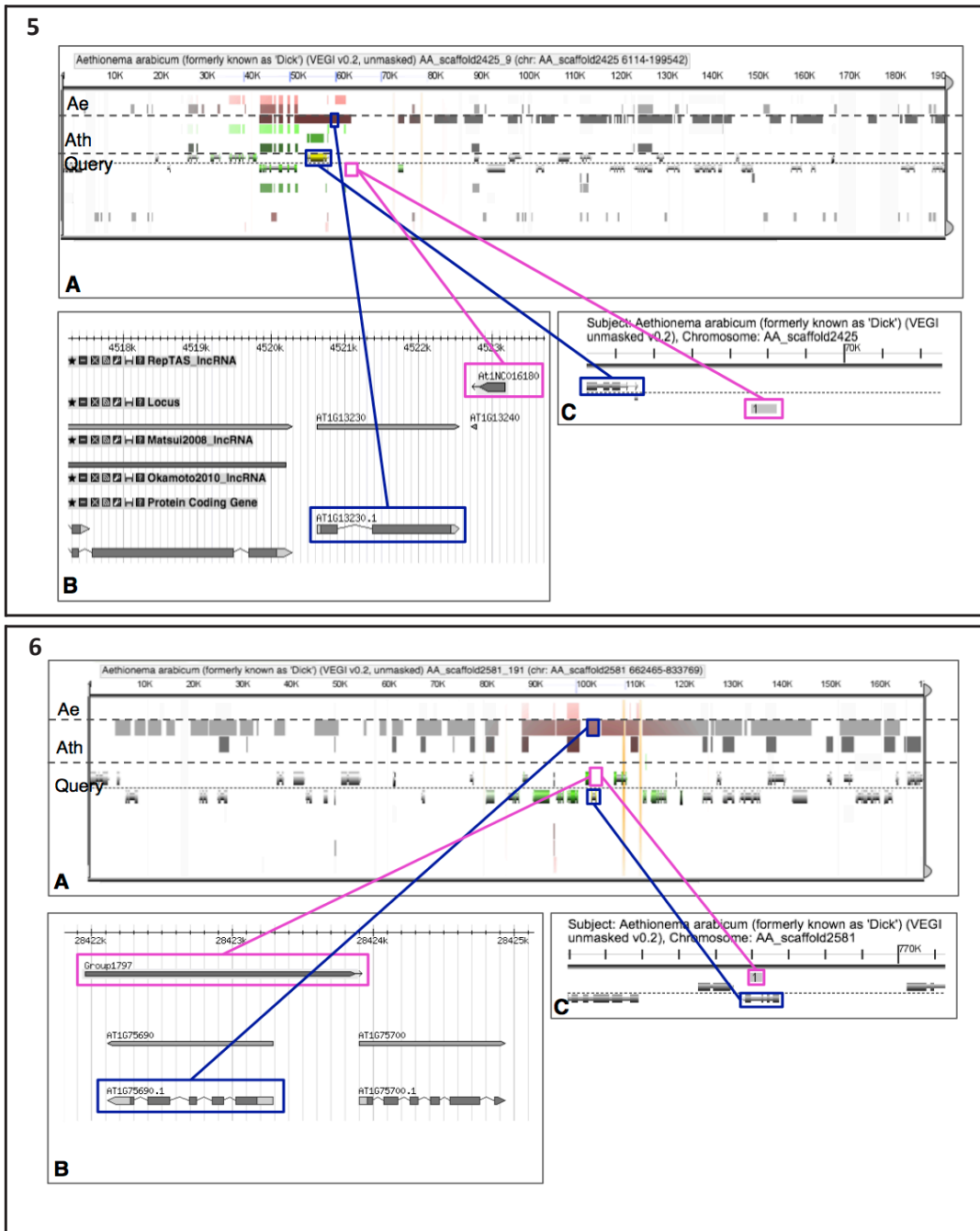


2

Suppl. Fig. 2 Pipeline to assess the LncRNAs specific to *Aethionemeae* or Cleomaceae.

2



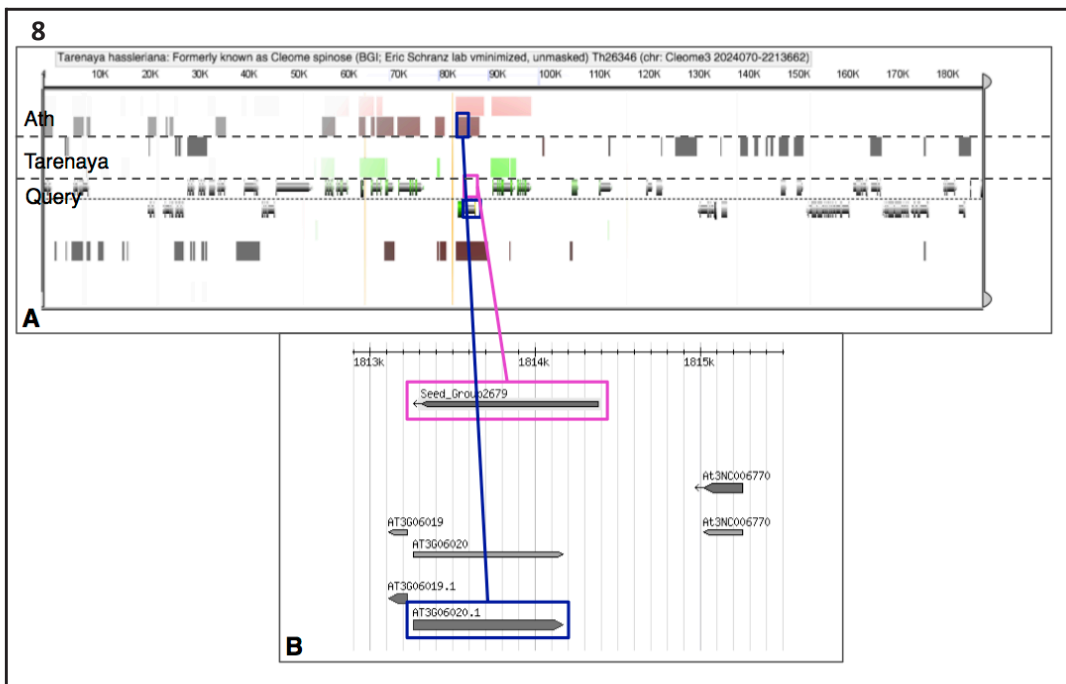
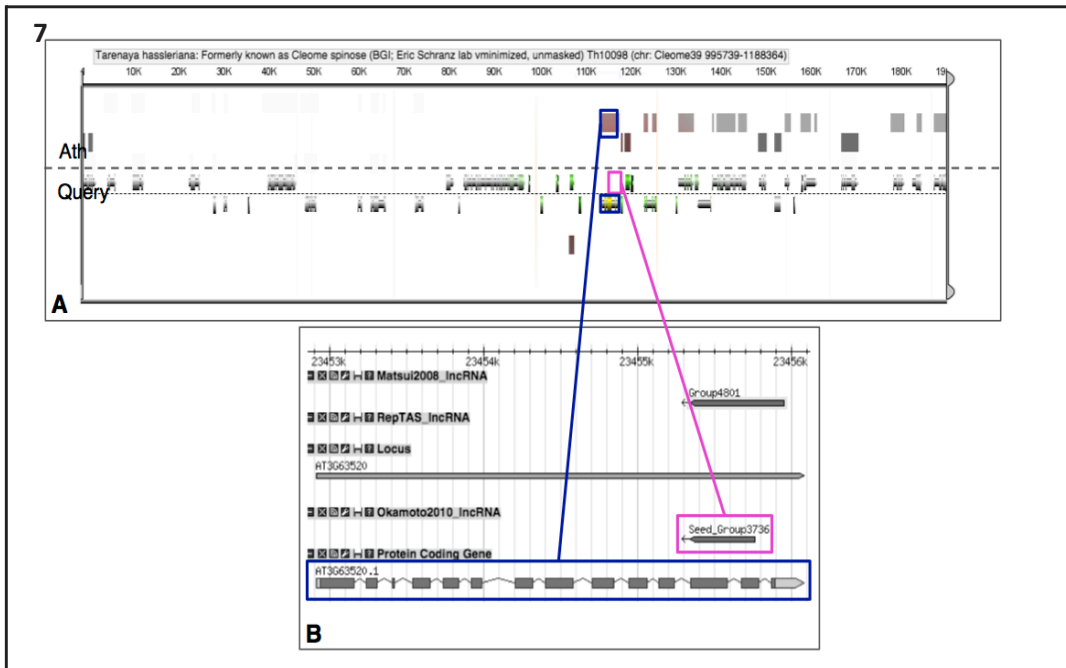


2

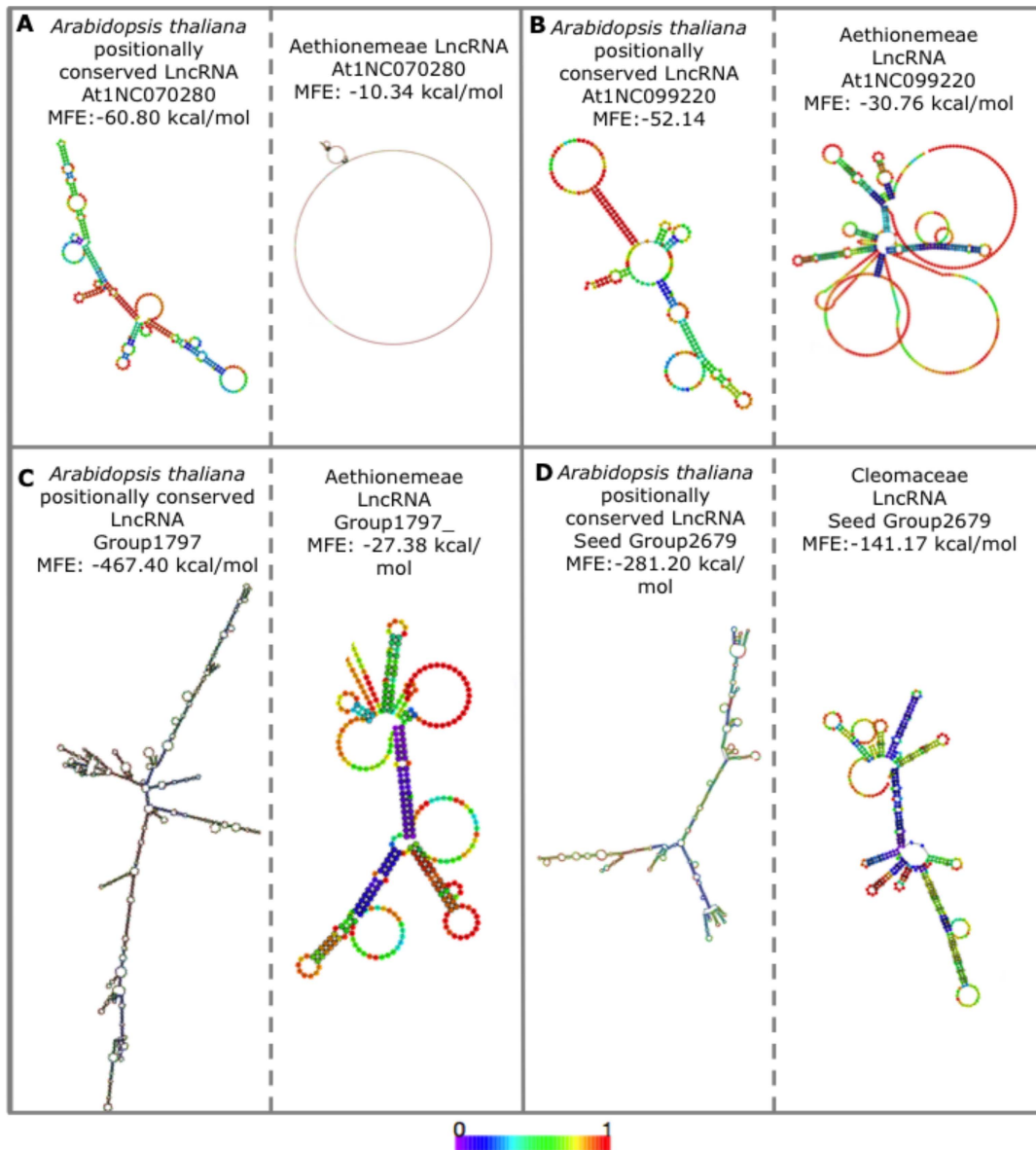
**Suppl. Fig. 3-6** Analyses of collinearity and positional conservation of sequentially diversified Aethionemeae LncRNAs. (A) Screenshot from GeVo. GeVo calculates the collinearity of a query sequence with the genome of a subject organism. The query here is the nearest protein coding gene of *Ae. arabicum* shown in B, the subjects are *Ae. arabicum* and *A. thaliana*. The position of the positionally conserved LncRNA is shown with a pink box, while the protein coding genes of *A. thaliana* and *Ae. arabicum* are shown with blue boxes. (B) Screenshot from the PIncDB website, shown are the *Arabidopsis thaliana* LncRNA (pink) and its nearest protein coding gene (blue). (C) Screenshot from the CoGe Blast HSP. Pink is the *Aethionema arabicum* transcript along the *Ae. arabicum* genome. Blue is the nearest *Ae. arabicum* protein coding gene.

Chapter 2

2

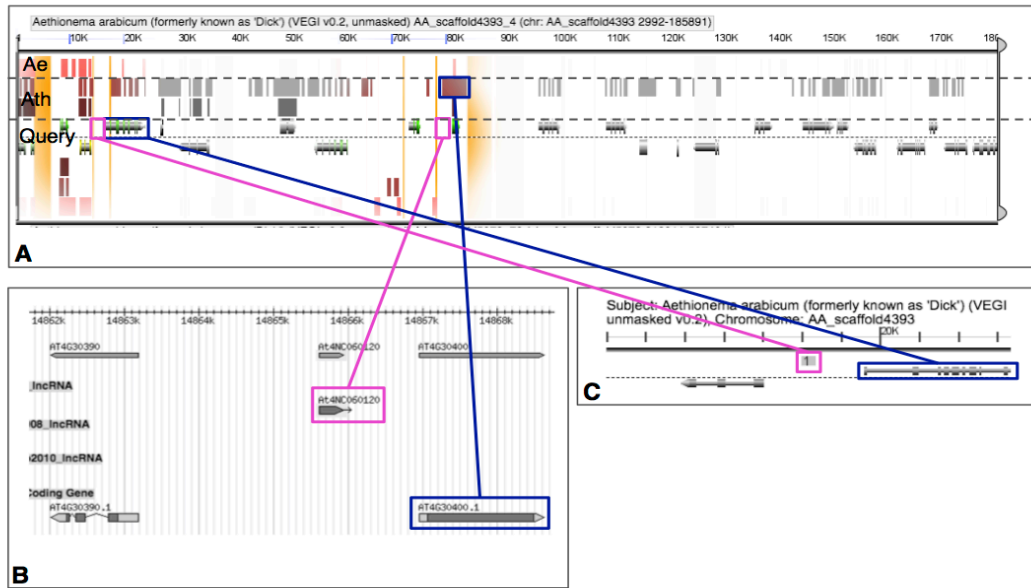






**Suppl. Fig. 10** Secondary structures and Minimum Free Energy (MFE) of sequence and/or positionally conserved LncRNAs. (A) LncRNAs that have both sequence conservation and positional conservation between *Arabidopsis* (left) and *Aethionema* (right) (B) LncRNAs that have only positional conservation between *Arabidopsis* (left) and *Aethionema* (right) (C) LncRNAs that have both sequence conservation and positional conservation between *Arabidopsis* (left) and *Tarenaya* (right) (D) LncRNAs that have only positional conservation between *Arabidopsis* (left) and *Tarenaya* (right) (E). The coloured bar below shows the basepairing probability for every structure.

Long non-coding RNAs in the Brassicaceae and Cleomaceae



2

**Suppl. Fig. 11** Example of an analysis of collinearity and no positional conservation of a sequentially conserved LncRNA. The example is a LncRNA conserved at the sequence level within the Brassicaceae. A) Screenshot from the PLncDB website, shown are the *Arabidopsis thaliana* LncRNA (green) and its nearest protein coding gene (blue). B) Screenshot from the CoGe Blast HSP. Green is the *Aethionema arabicum* transcript along the *Ae. arabicum* genome. Blue is the nearest *Ae. arabicum* protein coding gene. C) Screenshot from GeVo. GeVo calculates the collinearity of a query sequence with the genome of a subject organism. The query here is the nearest protein coding gene of *Ae. arabicum* shown in B, the subjects are *Ae. arabicum* and *A. thaliana*. The query here shows two collinear regions in *A. thaliana*. The position of LncRNA is shown with a green box, while the protein coding genes of *A. thaliana* and *Ae. arabicum* are shown with blue boxes. These SynFind and GeVo analyses can be redone with the following link: <https://genomeevolution.org/r/fmqj>







## Chapter 3

### **Anatolian origins and diversification of *Aethionema*, the sister lineage of the core Brassicaceae**

**Setareh Mohammadin<sup>1</sup>, Kim Peterse<sup>1</sup>, Sara J. van de Kerke<sup>1</sup>,  
Lars W. Chatrou<sup>1</sup>, Ali A. Dönmez<sup>2</sup>, Klaus Mummenhoff<sup>3</sup>, J. Chris Pires<sup>4</sup>,  
Patrick P. Edger<sup>5</sup>, Ihsan A. Al-Shehbaz<sup>6</sup>, M. Eric Schranz<sup>1</sup>**

American Journal of Botany, *in revision*

---

<sup>1</sup> Biosystematics, Plant Science Group, Wageningen University and Research, Wageningen, The Netherlands

<sup>2</sup> Department of Botany, Faculty of Science, Hacettepe University, Ankara, Turkey

<sup>3</sup> Department of Biology, Botany, University of Osnabrück, Osnabrück, Germany

<sup>4</sup> Division of Biological Sciences, University of Missouri, Columbia, USA

<sup>5</sup> Department of Horticulture, Michigan State University, USA

<sup>6</sup> Missouri Botanical Garden, St. Louis, MO 63166-0299, USA

## Abstract

**Premise of the study** The Irano-Turanian region harbours three biodiversity hotspots and ~25% of Brassicaceae species are endemic to the region. *Aethionema*, the sister lineage to the core Brassicaceae, contains ~61 species and occurs mainly in the Irano-Turanian region. The evolutionary important position of *Aethionema* makes it an ideal reference for broader comparative genetics and genomics. To better understand the evolution of *Aethionema*, and to bring us a step closer to understand crucifer evolution, a time calibrated phylogenetic tree and biogeographical history of the genus is needed.

**Methods** Seventy-six plastome coding regions and nuclear rDNA genes derived mainly from herbarium material, covering 75% of all *Aethionema* species, were used to resolve a time-calibrated phylogenetic tree of *Aethionema*. The different clades were characterised based on four morphological characters. Historical biogeographical analyses recovered the ancestral area of *Aethionema*.

**Key Results** The phylogenetic trees presented of *Aethionema* have three main well-supported resolved clades. The ancestral area reconstruction and divergence-time estimates are consistent with major dispersal events during the Pliocene from the Anatolian Diagonal.

**Conclusion** We find that most *Aethionema* lineages originated along the Anatolian Diagonal, a floristic bridge connecting the east to the west, during the Pliocene. The dispersal of *Aethionema* correlates with the local geological events, such as the uplift of the Anatolian and Iranian plateaus and the formation of the major mountain ranges of the Irano-Turanian region. Knowing the paleo-ecological context of the evolution of *Aethionema* facilitates our broader understanding for trait evolution and species diversification across the Brassicaceae.

## Background

The Irano-Turanian region harbours three main global biodiversity hotspots (Manafzadeh *et al.* 2016, Micó *et al.* 2009, Takhtajan 1986). With ~900 species, this region is also one of world's hotspots for Brassicaceae diversity (Koch & Kiefer 2006). Within the Western Irano-Turanian region Davis (1971) defined a floristic break called the Anatolian Diagonal. The diagonal runs from the rain-shadow of the Pontic mountains going southwest towards the Anti-Taurus (Davis 1971). The species diversity seen in the Irano-Turanian region and along the Anatolian Diagonal can be correlated with past tectonic events. The formation of the Alborz, Zagros and Kopeh-Dagh mountain ranges in Iran and the Taurus and Pontic mountain ranges in Turkey are all part of the Alpine orogeny, due to collisions of the African and Arabian plates with the Eurasian plate (Karl & Koch 2013, Davis 1971). In particular the collision of the Arabian plate with the Eurasian plate during the Miocene accelerated the uplift of the Alborz, Zagros and Kopeh-Dagh mountains and the lifting of the Iranian and Anatolian plateaus (Manafzadeh *et al.* 2016). The formation and uplift of these areas caused an aridification of the Irano-Turanian region. In addition the Irano-Turanian region became a melting pot of the climatic features of its surrounding regions (Djamali *et al.* 2012, Manafzadeh *et al.* 2016). This makes the Irano-Turanian region a place where species borders cross each other, allowing many species to co-occur and hybridize.

The plant family Brassicaceae (Cruciferae), or mustard family, contains several important vegetable crops (cabbage, broccoli, cauliflower, turnip, etc.) and several model plant species, among which *Arabidopsis thaliana* (L.) Heynh. is the best known. The Brassicaceae phylogeny is clearly split into a larger core clade with 324 genera and 3680 species (Al-Shehbaz, 2012) and a much smaller sister clade, containing only the genus *Aethionema* W.T. Aiton (61 species, The Plants List, 2013, Beilstein *et al.* 2006, 2008, Couvreur *et al.* 2010 Franzke *et al.* 2009, Huang *et al.* 2015). The divergence between *Aethionema* and the Brassicaceae core group occurred sometime during the Eocene, between 32-54 millions years ago (Mya) (Huang *et al.* 2016, Hohmann *et al.* 2015, Edger *et al.* 2015, Beilstein *et al.* 2010). Although the ages in Beilstein *et al.* (2010) and Edger *et al.* (2015) represent the outer limits for the divergence time, their confidence intervals overlap with the mentioned late Eocene divergence of the other studies (Hohmann *et al.* 2015, Huang *et al.* 2016). A more recent study (Cardinal-McTeague *et al.* 2016) suggests that the most recent common ancestor of all extant Brassicaceae originated somewhere in the broad geographical region of Europe, the Mediterranean Africa, Saharo-Arabian and/or Irano-Turanian regions during the middle Eocene (~43 Mya). Thus, more information about the origin of *Aethionema* could bring us a step closer in understanding the still unclear origin of the Brassicaceae family.

*Aethionema* occurs mainly in the western Irano-Turanian region (Hedge & Davis, 1965; Manafzadeh *et al.*, 2016), the hypothesized but not proven centre of origin for Brassicaceae (Al-Shehbaz *et al.* 2006, Couvreur *et al.* 2010, Franzke *et al.* 2011, Warwick *et al.* 2010, Hedge 1976). *Aethionema* species have ovate to linear leaves, are mainly perennial (though five species are annual) have mainly angustiseptate (flattened parallel to septum) dehiscent fruits (though some species also have indehiscent fruits making them heterocarpic) and some species produce spines (Lenser *et al.*, 2016; Prantl, 1891). The base chromosome number of *Aethionema* is likely seven or eight, although this can go up to  $x=24$  due to polyploidy events (Warwick & Al-Shehbaz 2006). The comparison of the now available genome of *Aethionema arabicum* (L.) Andr. ex DC. (Haudry *et al.* 2013) to the genomes

of the Brassicaceae core group showed that *Aethionema* and the rest of the Brassicaceae share the At-alpha whole genome duplication (Edger *et al.* 2015), have methionine-derived glucosinolates (Hofberger *et al.* 2013) and positionally conserved long non-coding RNAs and other conserved non-coding sequences (Mohammadin *et al.*, 2015, Haudry *et al.*, 2013). Hence, *Aethionema* shares the known cruciferous synapomorphic traits with the Brassicaceae core group. The evolutionary important position of *Aethionema* facilitates its use in a comparative framework to understand trait evolution and speciation across the Brassicaceae.

The phylogenetic position of the genus *Aethionema* historically has been uncertain. It was assigned to the tribe/sub-tribe of Sinapeae/ Cochleariinae (Prantl, 1891), Lepidieae/ Iberidinae (Hayek 1911), and Lepidieae/Thlaspidinae (Schulz 1936, Janchen 1942). Based on the chloroplast *rbcl* gene, Price *et al.* (1994) were the first to infer the sister group relationship of *Aethionema* with the core Brassicaceae. This phylogenetic position has been corroborated by additional studies (Bailey *et al.*, 2006; Beilstein *et al.*, 2006, 2008; Franzke *et al.*, 2009; Couvreur *et al.*, 2010; Warwick *et al.*, 2010). However, five traditionally recognized *Aethionema* species have been synonymized into the Brassicaceae core-group genus *Noccaea*: *N. trinervia*, *N. oppositifolia*, *N. iberidea*, *N. rotundifolia* and *N. apterocarpa* (Al-Shehbaz 2012). These species were moved based on *Aethionema* specific morphological synapomorphies (Al-Shehbaz 2012). Al-Shehbaz *et al.* (2006) were the first to recognize the tribe Aethionemeae, consisting of the genera *Aethionema* and *Moriera*. However, as Prantl (1891) and Komarov (1939) already suggested, *Moriera* is now a synonym of *Aethionema* (Al-Shehbaz 2012) making Aethionemeae a unigeneric tribe.

Here we present a comprehensive phylogenomic framework of *Aethionema* based on whole plastome and nuclear ribosomal DNA sequences generated mainly from herbarium specimens. The mixture of a high species coverage with a large number of molecular markers allows us to study the following objectives: i) do the *Aethionema* clades correlate with morphological characteristics, e.g. leaf shape and heterocarpism? ii) What is the divergence time of *Aethionema*? iii) Does the biogeographical distribution and history of *Aethionema* correlate with geological and climatic conditions of the Irano-Turanian region? Knowing the phylogenetic and biogeographical history of *Aethionema* species divergences adds a piece to the puzzle to understand the paleo-ecological context of crucifer diversification.

## Materials and Methods

### Taxon sampling, sequencing and assembly

We sampled fifty different *Aethionema* species (*A. turcicum* and *A. transhyrcanum* consisted of two samples each), five species of *Noccaea* (*N. trinervia*, *N. oppositifolia*, *N. iberidea*, *N. rotundifolia* and *N. apterocarpa*) formerly classified in *Aethionema*, and an outgroup species *Cleome droserifolia*, from the sister family Cleomaceae. Thirty-nine of these fifty-eight accessions were sampled from herbaria with the oldest one, *A. membranaceum*, collected in 1867 and the youngest in 2008 (Suppl. Table 2). Nineteen out of the fifty-eight samples were silica-dried samples collected in the field (Suppl. Table 2).

DNA isolation and library preparation was conducted in two batches named as ‘Copenhagen’ and ‘Missouri’, based on the location where the samples were sequenced (Suppl. Table 2).

DNA isolation, library preparation and sequencing for the ‘Copenhagen’ samples were done according to Bakker *et al.* (2015). DNA isolation and library preparation for the ‘Missouri’ samples was conducted as follows. Leaves collected from the Missouri Botanical Garden Herbarium were ground in liquid nitrogen and kept at -20°C until further use. DNA was isolated with the Qiagen DNeasy Plant Mini Kit (Qiagen, Hilden, Germany). Quality and quantity checks were done on a NanoDrop 1000<sup>®</sup> spectrophotometer (Thermo Fisher Scientific, Wilmington, DE, USA). The DNA was eluted with 7% demi-water, sheared using sonication in time laps of 7.5 minutes alternating every 30 sec. on and off and concentrated with the Qiagen Qiaquick PCR purification microcentrifuge kit (Qiagen, Hilden, Germany). Library preparation was done with the Illumina TruSeq<sup>™</sup> DNA Sample Preparation kit (Illumina, San Diego, CA, USA) using the Low-Throughput protocol following manufacturer’s protocol with an insert size varying between 300-500 bp. Final quality check before sequencing was done on a Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA). The samples were sequenced on the Illumina HiSeq-2000 with paired-end 100bp reads at the sequencing core of the University of Missouri in 2011.

Both sample batches were assembled using IOGA v1.6 (Bakker *et al.* 2015). IOGA is an iterative assembly pipeline that uses a reference file to assemble a genomic component. To assemble the plastid genomes (plastomes), we used the same reference plastomes as in Bakker *et al.* (2015). For the nuclear rDNA regions, we used the *Brassica rapa* complete rDNA downloaded from NCBI genbank (gb code: KM538956.1). Samples with fewer than 92,784 reads were discarded. Samples with more than 26 million reads were subsampled before assembly using seqtk (seed=100, subsample size =1,000,000). Suppl. Table 2 contains the assembly summaries. *Aethionema diastrophis*, *A. kopetdaghi*, *A. caespitosum*, *A. marashicum* and *A. papillosum* did not have enough read coverage for assembly and were hence discarded. After assembly, there were a total of 46 *Aethionema* species, the five *Noccaea* species and the outgroup species *Cleome droserifolia*.

### Annotations, alignment and data matrix construction

The chloroplasts were annotated using the *Arabidopsis thaliana* chloroplast as a reference. This resulted in seventy-one protein coding genes, 20 tRNAs and 3 rRNAs for every accession. For downstream analyses we used only the protein coding genes and excluded the introns having a total of seventy-six coding sequences. In addition to our samples, we added full plastome data of twenty-seven core Brassicaceae species from Hohmann *et al.* (2015). These twenty-seven additions are representative species from Brassicaceae lineage I, lineage II and expanded lineage II according to Beilstein *et al.* (2006), lineage III is not represented. It is unlikely that including lineage III would have altered the results for our study focused on *Aethionema*. To be consistent throughout all our analyses, we annotated the core group twenty-seven chloroplasts in the same way as our own samples. Nucleotide alignments were made for each gene with the MAFFT (Katoh & Standley 2013) plugin of Geneious v8.8 (Kearse *et al.* 2012). We made one final data matrix, the chloro-matrix, by concatenating the individual alignments with Sequence Matrix (Vaidya *et al.* 2011). The chloro-matrix consisted of eighty terminals with 59,360 nucleotides (10,531 variable and 5,959 parsimony informative sites). Suppl. Table 2 shows for every terminal which genes are present and their length. The used chloro-matrix alignment is given in Suppl. Dataset 1.

We annotated the nuclear rDNA coding regions (18S, 5.8S and 26S) and ITS1 and ITS2 with *Brassica rapa* as a reference. BlastN (Altschul *et al.* 1990), with E-value  $\leq e^{10}$  and query identity  $\geq 50\%$ , was used to check the newly generated rDNA sequences for bacterial or fungal hits. None were found. The 5.8S, ITS1 and ITS2 sequences of 24 Brassicaceae core-group species, all present in the chloro-matrix, were downloaded from NCBI (Suppl. Table 2). The alignment and concatenation of the rDNA-matrix was done similarly as for the chloro-matrix. The data we generated included all present rDNA coding regions and ITS1 and ITS2, while the core group consisted only of the 5.8S, ITS1 and ITS2 sequences (Suppl. Table 2). The final rDNA-matrix consisted of 5,857 nucleotides (206 variable and 443 parsimony informative sites) and 72 terminals. Suppl. Dataset 2 contains the used rDNA alignment.

### Phylogenetic analyses

Phylogenetic trees for the chloro-matrix were inferred under maximum likelihood using RaxML v 8.2.4 (GTR+GAMMA, random seed and 1000 bootstrap pseudoreplicates) on the CIPRES science gateway (Stamatakis 2014, Miller *et al.* 2010). To assess whether partitioning a large dataset with 59,360 nucleotides has an effect on the support of the backbone of our phylogenies we used biologically intuitive partitioning schemes as well as one objective assessed by a programme. Four different partitioning schemes were compared: unpartitioned, partitioned by codon position (3 partitions), partitioned by coding region (76 partitions), and one assigned by PartitionFinder (Lanfear *et al.* 2012). PartitionFinder v1.1.1 was used to find a partition scheme among the 76 coding regions and their codon positions using the GTR models of RaxML for models of evolution and the Bayesian information criterion (BIC) for model selection with a greedy search scheme. The different partition schemes were compared using the Akaike Information Criterion (AIC) following Barrett *et al.* (2016). Maximum Parsimony (MP) analyses were done in PAUP\* 4.0a150 (Swofford 2002) with tree bisection reconnection, 75 random addition sequence replicates, 1000 jackknife replicates and 37% deletion probability. A majority rule consensus tree was made with a 50% cut-off of all the saved trees.

For the rDNA-matrix we used the same approach as for the chloro-matrix. The partitioning schemes here were: unpartitioned, partitioned per gene, ITS regions vs. coding genes and a PartitionFinder scheme. Maximum likelihood analyses were run as described above for these partitioning schemes. In addition to a RaxML analyses, we used MrBayes 3.2.6 (Huelsenbeck & Ronquist 2001, Ronquist & Huelsenbeck 2003) for a Bayesian inference of our phylogenies (250 million generations, on four chains, sampled every 250,000<sup>th</sup> generation, temp=0.05 and diagnfreq=5000).

To assess whether the *Aethionema* clades could be characterized by morphological characters, we scored leaf shape (ovate vs. linear), fruit morph (dehiscent vs. heterocarpic), presence or absence of spines and plant duration (annual vs. perennial) for the forty-six *Aethionema* species using Flora Iranica, Flora of Turkey, publications of recently published species and herbarium sheets (Rechinger, 1968; Hedge, 1965, 1968; Pavlova, 2007; Ertuğrul & Beyazoğlu, 1996).

### Divergence date estimates

We inferred a time calibrated phylogenetic tree using Bayesian methods to understand the timing of speciation of *Aethionema* using the chloro-matrix. We used an uncorrelated relaxed molecular clock to account for rate variability among lineages as implemented in BEAST v1.8.3 (Drummond et al. 2006, 2012). BEAUti v1.8.3 (Drummond et al. 2012) was used to create a XML file. Secondary calibration of the (lower and upper bound of 47.8-60.6 Mya) crown-node age of Brassicaceae was set as a uniform prior, including confidence intervals, following the most recently published crown node age of the Brassicaceae (Cardinal-McTeague et al. 2016) overlapping with other estimates of the known crown-node age (Hohmann et al. 2015, Edger et al. 2015, Beilstein et al. 2010, Couvreur et al. 2010, Huang et al. 2016). Monophyly of the Brassicaceae family was constrained and the mean root height of the random starting tree was set to 53 Mya. The uncorrelated relaxed clock model had a log-normal distribution substitution rate prior and the Yule speciation model as tree prior. We started seven Markov Chain Monte Carlo (MCMC) analyses of 150 million generations each, sampling every 1000 steps. Four out of these seven analyses had a posterior and prior ESS>200 as assessed in Tracer v1.6 (Rambaut et al. 2014). Trees from these four runs were combined using LogCombiner v1.8.3 and summarized in TreeAnnotator v1.8.3 with a 10% burn-in of 60,000 trees to get a maximum clade credibility tree with mean heights.

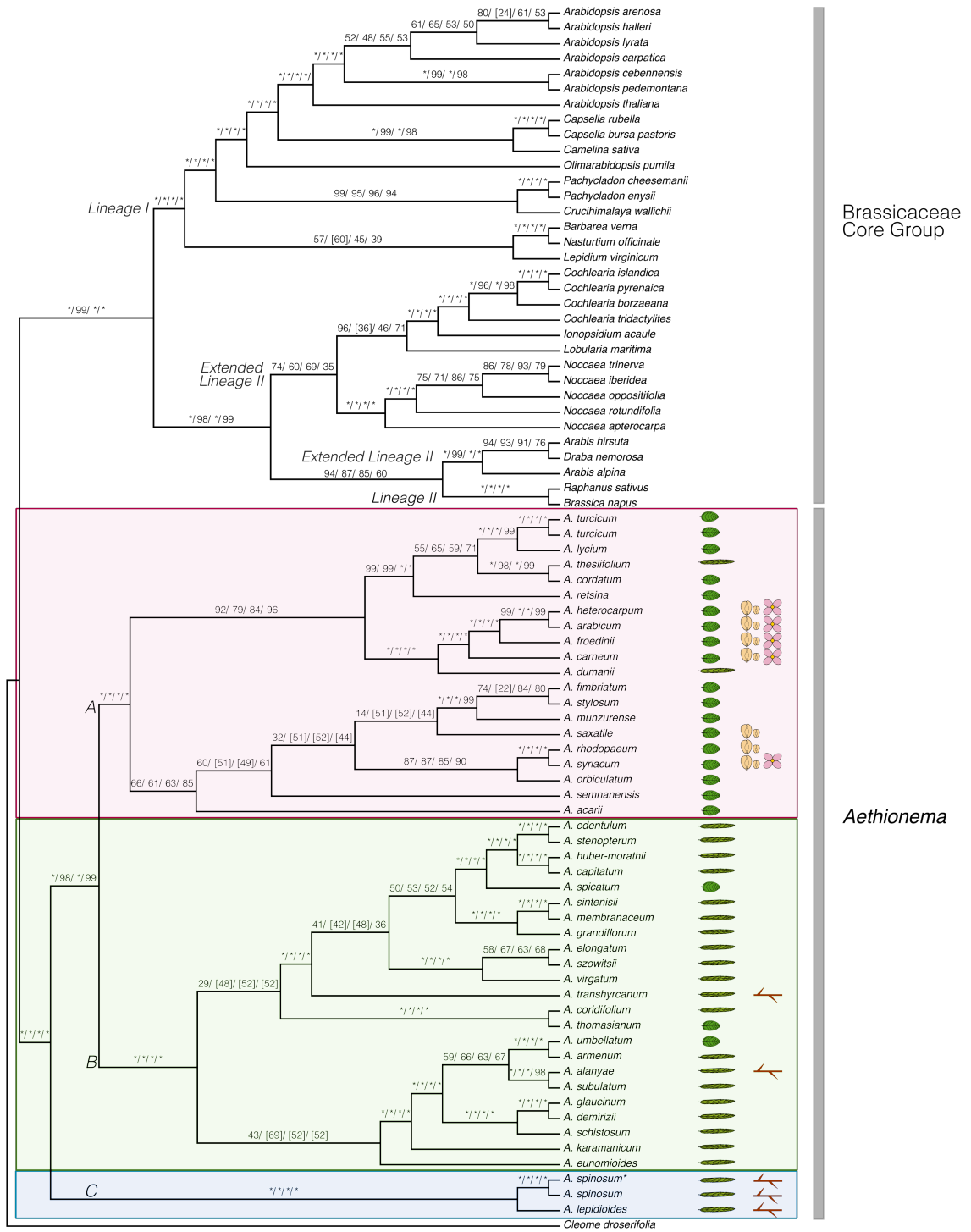
### Historical biogeography reconstruction

To understand the geographical distribution and historical patterns of *Aethionema*, we conducted an ancestral area reconstruction. The digitized herbarium sheets available from the Global Biodiversity Information Facility (GBIF 2012) were used to estimate the geographical distribution of all used *Aethionema* species. To reconstruct the ancestral areas of *Aethionema*, we used specimens belonging to *Aethionema sensu stricto* (44 species), hence excluding the above-mentioned *Noccaea* species. Many *Aethionema* have been collected without GPS data, thus the location of collection and the coordinates were assessed by hand directly from the herbarium sheets (Suppl. Fig. 1, final coordinates used per species: Suppl. Table 2). Forty-one percent of the used herbarium sheets have been curated in the last fifteen years (Al-Shehbaz, Dönmez), reducing misidentifications (Goodwin et al. 2015).

Our data showed that *Aethionema* species are distributed from the Iberian Peninsula in the west to Afghanistan and Turkmenistan in the east, and from Switzerland in the north to Tunisia in the south. We made 14 area delimitations based on species distribution and geography following the floristic regions of Takhtajan (1986) and the eco-regions of Olson et al. (2001). These regions followed geological barriers (mountains), containing at least two *Aethionema* species following as much as possible the spatial boundaries of the species. The regions were as follows (Fig. 2): A) western Mediterranean, B) north eastern Mediterranean, C) central Turkey, D) west-central Taurus and Cyprus, E) the Anti-Taurus, F) northern Mesopotamia, G) eastern Pontic mountain range, H) lesser Caucasus area, I) (anti-) Lebanon mountain range, J) Israel mountain ranges, K) coastal areas of the Arabian plate and the northern Nubian shield, L) Alborz-Kopeh Dagh mountain range (Hyrcanian forests), M) Zagros mountains, and N) Hindu-Kush mountains.

Ancestral area reconstruction was done with S-DIVA and BBM in RASP v3.2 (Yu et al. 2015) with maximally three areas allowed for both methods. For both methods we used 20,000 randomly selected post burn-in trees from the combined Beast analyses and the tree

3



**Fig. 1** Cladogram based on the maximum likelihood analysis (GTR+GAMMA, unpartitioned) of the chloro-matrix (76 plastome coding regions). Numbers along the branches are the bootstrap values for different partitioning schemes: unpartitioned/ per codon partition/ PartitionFinder/ partition per gene. \*=100% bootstrap value. Bootstrap values in brackets = alternative relationship for that branch. Icons along the *Aethionema* clade show: linear/ovate leaves, heterocarpic species (two different sized fruits), annual species (pink flowers) and species with spines (spiny branch).



provided by combined tree from our divergence time analysis as a condensed tree. For BBM analyses we ran 30 million generations on four chains, sampling every 10,000<sup>th</sup> generation, F81 as state frequency and a gamma distribution for the among site rate variation. To be certain that species with more than one representative in our phylogenetic tree would not bias the ancestral area reconstruction we reduced each terminal to one representative. Hence we excluded the duplicate *Aethionema turcicum* and *A. spinosum*\* from our RASP analyses. Our focus here was to understand the dispersal and vicariance processes and the ancestral area of *Aethionema*, hence we excluded the core group from our historical biogeographical analyses.

## Results

### Plastome and nuclear rDNA assemblies

DNA and library preparation was done from herbaria and silica-dried materials on 58 samples from 52 *Aethionema* specimens (50 different species), five *Noccaea* species and the outgroup species *Cleome droserifolia* from the family Cleomaceae. The sequencing was successful for 51 specimens, of which 45 were different *Aethionema* species (75% species coverage). The number of reads over all samples ranged from 92,784-103,154,570 before read trimming and quality checks. Trimming and quality checks are incorporated in the IOGA pipeline (Bakker *et al.* 2015) using Trimmomatic (Bolger *et al.* 2014). The assembly length of the chloroplasts had a range of 123,681-228,766bp with a read coverage ranging between 8.3-627.2x. The average read coverage of 112.18x corresponds to 89% of the *Arabidopsis thaliana* chloroplast (154,478bp as present in CoGe (Freeling *et al.* 2008)). The percentage of chloroplast reads present in the unfiltered dataset ranges from 0.99% to 26.71%. Average GC content was 37.18% (SD±4.78%), a bit higher than the GC content of the *A. thaliana* chloroplast (36.3%).

We assembled the nuclear ribosomal DNA (rDNA) genes and Internal Transcribed Spacer (ITS) regions for 42 *Aethionema* species, the five *Noccaea* species and *C. droserifolia*. The assembly length ranged from 6,688bp to 28,006bp with a read coverage between 20.1-1301.4x. The length of the assembled rDNA sequences was on average 137% compared to the *Brassica rapa* reference. This is probably due to the iterative character of the assembly pipeline that keeps assembling reads and expanding the contigs. However, as we retrieved the individual rDNA regions from our assemblies the iterative character of the IOGA-pipeline did not have consequence for phylogenetic analyses or ancestral area reconstruction.

### Phylogenetic reconstruction of *Aethionema*

To assess whether partitioning 59,360 markers (chloro-matrix) had an effect on the resolution of the backbone of the *Aethionema* phylogeny, we used different partition models varying between no partition to a per gene partition (Table 1). The AIC values of the maximum likelihood analyses showed the lowest score for the model found by PartitionFinder for the chloro-matrix, but were the lowest for the per-codon partition for the rDNA-matrix (Table 1). MrBayes analyses of the rDNA-matrix did not converge for two out of the five partitions; hence in addition to the RaxML analyses of the rDNA-matrix we took only the MrBayes unpartitioned, ITS vs. coding genes partition and per gene partitioning analyses into account (Suppl. Fig. 7). The topology of the Brassicaceae core group in our study was the same as

in Hohmann *et al.* (2015). The backbone of the *Aethionema* chloroplast phylogeny was well supported by all partition models and by the parsimony analysis (Suppl. Fig. 2A) for both the rDNA-matrix (Suppl. Fig. 3) as well as the chloro-matrix (Fig. 1, Suppl. Fig. 2B for branch lengths) supporting the three clades A, B and C that Lenser *et al.* (2016) found with only two chloroplast markers. Hence, partitioning our large dataset does not have an effect on the already well-supported nodes. One conflicting species is *A. transhyrcanum* occurring in clade B as well in clade C. However, the herbarium sheets of these samples showed that the C-clade *A. transhyrcanum* used to be called *A. spinosum*, hence we assume here that the sheet has been misidentified and regard this sheet as an *A. spinosum* specimen and we denote it consistently as *A. spinosum*\*. The A, B, C clades can be roughly distinguished by the shape of their leaves and with the presence or absence of spines (Fig. 1). The *Aethionema* A clade includes the annuals as well as the heterocarpic species and consists mainly of ovate leaf shaped species. The B clade consists mainly of species with linear shaped leaves. All the C clade species have spines. In contrast to Lenser *et al.* (2016), our data contained all known heterocarpic and spiny species and consisted of many more markers. Hence our analyses give high support values to the independent origin of the annual, heterocarpic and spiny species.

### Divergence date estimates

*Aethionema* and the core Brassicaceae diverged in the Paleogene with the middle Eocene (~48 Mya, 95% HPD = 58.94-37.5 Mya) as its most probable date of origin (Figure S4). The diversification of the A, B clades from the C clade of *Aethionema* have its most probable origin at the beginning of the Miocene (~24 Mya, 95% HPD=34.58-14.61). This is in accordance with the diversification of the lineages within the Brassicaceae core group

**Table 1.** Partition models used in the maximum likelihood analyses for the chloro and rDNA matrix.

<b>Chloro-matrix</b>	GTR model Partitioning	No. Partitions <sup>a</sup>	log <sub>e</sub> L	Number of Free parameters	AIC
	Unpartitioned	1	-200558.61	166	401450.16
	Partitioned per codon	3	-199167.71	184	398704.57
	PartitionFinder	12	-191324.48	265	383181.35
	Gene partition	76	-190930.67	841	383567.54
<b>rDNA-matrix</b>	GTR model Partitioning	No. Partitions <sup>a</sup>	log <sub>e</sub> L	Number of Free parameters	AIC
	Unpartitioned	1	-17693.13	150	35694.20
	PartitionFinder	2	-16969.05	159	34265.02
	ITSvsCoding Partition	2	-16969.05	159	34265.02
	Gene partition	5	-16772.58	186	33929.44
	Partitioned per codon	11	-16593.58	240	33687.75

<sup>a</sup>) number of partitions used for a model.

Number of free parameters and the AIC values calculated after (Barrett *et al.*, 2016).

(Suppl. Fig. 4). The *Aethionema* A and B clade started diversifying approximately ~9 Mya (A clade 95% HPD = 12.94-6.91, B clade 95% HPD = 14.52-6.31), while the diversification of the C clade occurred approximately ~12 Mya (95% HPD = 20.62-5.92). Hence, the diversification of *Aethionema* species occurred in the Middle Miocene corresponding with orogeny of the mountain ranges in the western Irano-Turanian and the uplift of the Anatolian and Iranian plateaus (Manafzadeh *et al.* 2016).

### Ancestral area reconstruction

In total we used 864 coordinates of *Aethionema* and *Cleome droserifolia* specimens for the ancestral area reconstruction (Suppl. Table 2). In general, ancestral area reconstruction had low support (<50%), both with S-DIVA and BBM (Fig. 2). Fig. 2 shows for all nodes of interest the three regions with the highest probabilities. The RASP results are added as supplementary data (Suppl. Datasets 3 and 4). S-DIVA as well as BBM inferred the ancestral area of *Aethionema* clade A originated mainly from the northeastern Mediterranean (region B, S-DIVA 93.04%, BBM 24.92%), the combination of BC (the northeastern Mediterranean and central Turkey BBM 6.31%). Clade B was mainly inferred to originate from the Anti-Taurus mountains (BBM; E=4.61%) or from a combination of the eastern Pontic mountain range and the Anti-Taurus (S-DIVA; G=13.1%, E=7.22% and EG=6.54%). According to the S-DIVA analysis, the spiny *Aethionema* clade C does not have a distinct centre of origin, as all the combinations of the Anti-Taurus mountains, the eastern Pontic mountain range, the Hyrcanian forests, the Zagros Mountains and the Hindu-Kush mountains have the same probability of 6.67%. However, the BBM analysis inferred the eastern Pontic Mountain range and the Hyrcanian forest (G=15.84%, L=9.83% and GL=8.11%) as the origin of the C clade. The most recent common ancestor of *Aethionema* originated on the west side of the Anatolian diagonal (E and G) as shown by the BBM (E=8.1% and G=7.5%) and the S-DIVA analyses (E=47.14% and G=47.14%). The low probabilities and the lack of a coherent ancestral area reconstruction could be due to a complicated distribution pattern among the *Aethionema* clades (Fig. 2). To explain the origin of current distribution patterns, the historical biogeographical analysis inferred either 136 dispersal and 25 vicariance events (BBM analysis, Fig. 2) or 109 dispersal and 20 vicariance events (S-DIVA analysis) for *Aethionema*. There are 43 internal nodes in the *Aethionema* phylogeny (excluding the core group and the double terminals). Hence 32% (BBM) to 40% (S-DIVA) of the speciation events was followed by a dispersal event and probably allopatric speciation. The dispersal events peaked three or four times throughout the *Aethionema* evolutionary history. These peaks occur after the major geological events in the Irano-Turanian region (Fig. 2).

### Discussion

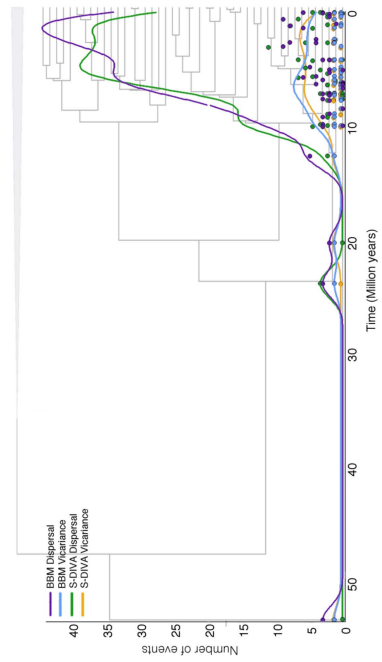
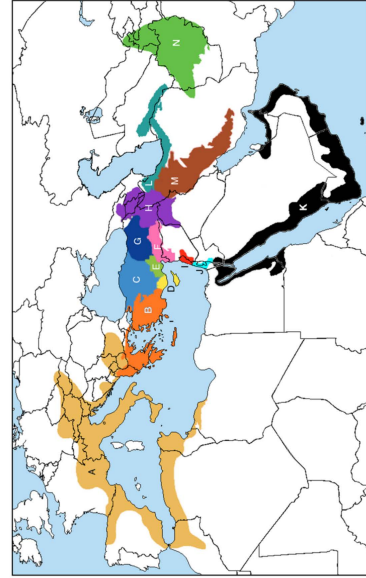
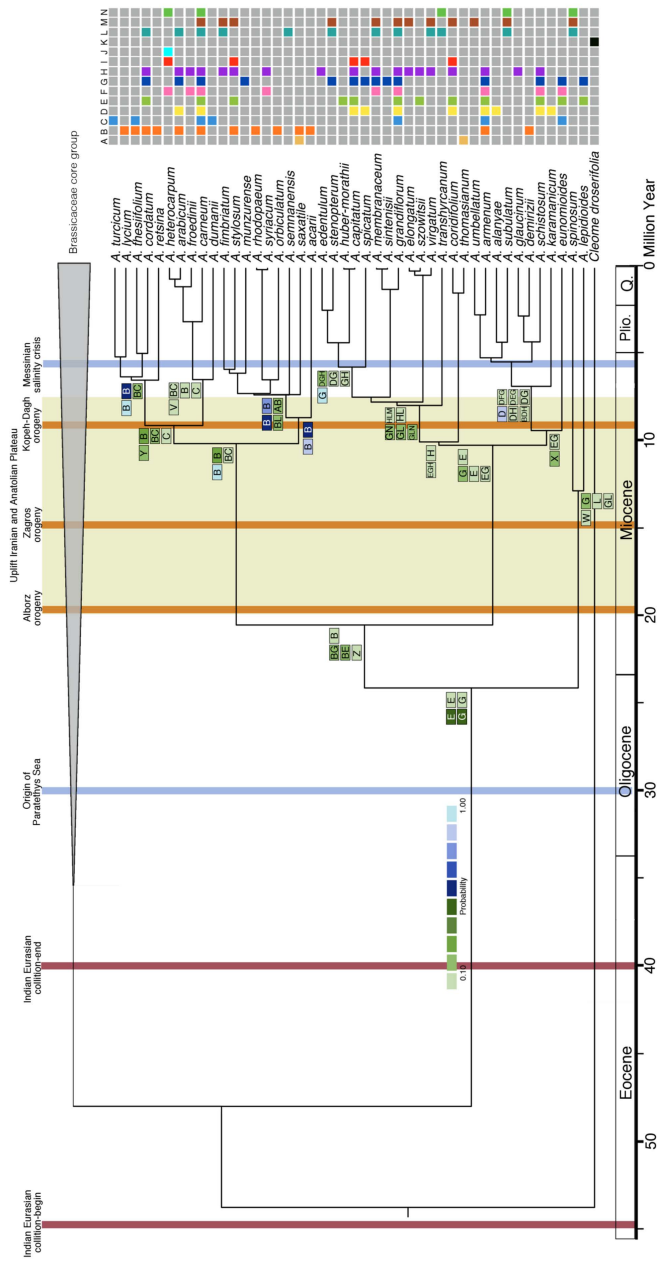
The Irano-Turanian region has been regarded as the potential centre of origin of the Brassicaceae family (Hedge 1976, Franzke *et al.* 2011) and is hotspot of species diversity for this economically important plant family. Nevertheless, there are only a few studies, like the present one, exploring biogeographic patterns and evolution of ecologically important characters of plant groups in this area (Özüdoğru *et al.* 2015 and references therein). Using 76 chloroplast markers and the nuclear rDNA we inferred a time-calibrated phylogeny showing that *Aethionema*, the sister of the Brassicaceae core group, has diversified along the Anatolian diagonal and dispersed through the Irano-Turanian region due to major geological events. *Aethionema* species likely dispersed from the Anatolian Diagonal into the

rest of the Western Irano-Turanian area. Their current distribution may have been caused by the closing of the connection of the Tethys Sea with the Indian Ocean, due to the full collision of the Arabian Plate into the Eurasian plate during the Miocene-Pliocene boundary (Davis 1971).

In addition to *Aethionema* the Irano-Turanian regions seems to harbour the origin of lineages from the Brassicaceae core. Mummenhoff *et al.* (2001) showed that Irano-Turanian and Mediterranean *Lepidium* species *s.str.* are never in a derived position in the *Lepidium* phylogeny. Although this does not exclude local extinctions of *Lepidium* species, they concluded that the genus *Lepidium* has its centre of origin in the Irano-Turanian and Mediterranean region attaining its current global distribution through migrations into North America via the Bering Land Bridge and into Australia via long-distance dispersal from western California, followed by hybridization with South African species (Mummenhoff *et al.* 2001, 2004; Koch and Kiefer 2006; Dierschke *et al.* 2009). Using one nuclear marker and one plastome spacer Özüdoğru *et al.* (2015) showed that the genus *Ricotia* likely originated from the Antalya region or the Anatolian Diagonal. The genus *Heldreichia* has its highest morphological leaf diversity in the south east of the Anatolian Diagonal (Parolly *et al.* 2010). With a haplotype analysis Ansell *et al.* (2011) showed that *Arabis alpina* had its origin and highest genetic diversity in Anatolia. Moreover, Ansell *et al.* (2011) showed that the Anatolian Diagonal might have functioned as bridge for the range expansion of *A. alpina*. Not only species and genera had their origin in the Irano-Turanian region. Karl & Koch (2013) showed that the diversification of the tribe Arabideae happened most likely in Anatolia and the Levantine coast. Hedge (1976) postulated that the most species rich region is probably also the cradle of origin for a lineage. However, this hypothesis does not always hold up. For example, the Andes mountain range contains the highest species diversity of the largest Brassicaceae genus *Draba*. However the centre of origin of *Draba* correlates with the region with the highest genetic diversity: the Irano-Turanian region (Jordon-Thaden 2009, Jordon-Thaden *et al.* 2013).

Our plastome and nuclear rDNA phylogenies provide independent evidence that the genus *Aethionema* can be split in three main clades (A, B and C). The three clades can roughly be distinguished morphologically based on leaf shape and/or the presence of spines. Our results agree with the maximum likelihood analysis of Lenser *et al.* (2016), based on only two chloroplast markers and fewer taxa, and resolve that heterocarpic infructescence, annuality and spine formation have independent origins. *Aethionema* clade A probably originated from the western Anatolian/eastern Mediterranean region, while clade B had an Anatolian Diagonal origin, similar to the origin of the *Aethionema* (Fig. 2). The north/east – south/west orientation of the diagonal provides the connection between the east-west

**Fig. 2 (Next page)** Ancestral area reconstruction of *Aethionema* along a geological time scale. Blocks on the right correspond with the occurrence for every terminal. Grey is used as a background, colour all other colours match with the areas in the inserted map. The three highest S-DIVA and BBM results are given for fifteen nodes with coloured blocks. The colour corresponds with the probability of reconstructed ancestral area (range given along ancestral *Aethionema* branch) and the letter within the block the representing area. Letters other than A-N denote the following: V= CH, BCN, BCG BCM, BCL, BCH, BCD, BC, C, BCF, CF, BCE; W= EGM, GM, EM, EMN, GLM, ELM, GMN, GLN, EGN, GN, EN, EL, ELN, EGL, GL; X= CDG, DEG, DFG, DG, DI, CDE, DEF; Y= BCF, BH, BCH, B, C, BF; Z=BDG, BDG. Geological events are shown with coloured bars along the graph. They have been adapted from Manafzadeh *et al.*, (2016). The graph on the lower left bottom shows the dispersal and vicariance events from the S-DIVA and BBM analyses.



3

directed mountains of the Mediterranean and the Caspian Sea (Ansell *et al.* 2011, Médail & Diadema 2009). The Anatolian Diagonal has been shown to serve as a barrier for some animal species (Gül 2013, Vamberger *et al.* 2013) or as a diversification spot for other groups (Stümpel *et al.* 2016). However, as Davis (1971) already suggested, the Anatolian Diagonal might have served as a land bridge between east and west for plants. This hypothesis has been recognized for the Brassicaceae genera *Lepidium*, *Draba* and *Arabis*, specially *Arabis alpina* (Mummenhoff *et al.* 2001; Jordon-Thaden 2009; Ansell *et al.* 2011; Karl & Koch 2013) and in this research for *Aethionema*.

The dispersal of *Aethionema* happened mainly during the Pleistocene when an ice sheet covered the European continent that did not reach into the Mediterranean and Irano-Turanian region (Ansell *et al.* 2011, Micó *et al.* 2009). The temperate climate in the Irano-Turanian region during the glacial periods might have made it possible for *Aethionema* species to disperse, as they tend to grow at high elevations and flower before the summer heat (Bibalani 2012). *Aethionema* dispersed and probably went through allopatric speciation cycles, including hybridization and polyploidization, after the newly formed mountain ranges caused topographical heterogeneity in the Irano-Turanian region. The *Aethionema* diversification coincided with the Arabideae diversification, the largest Brassicaceae tribe (Karl & Koch 2013). It has been hypothesized that the rapid radiation seen in the Brassicaceae might have been due to the evolution of novel traits (Schranz *et al.* 2012). The co-evolution of the Pirideae butterflies and Brassicaceae defence compounds (glucosinolates) against the Pirideae caterpillars is an example of a co-evolved novel trait (Edger *et al.* 2015). Our research puts the evolution of these novel traits in a time frame where local geological changes might have had a hand in the radiation of a plant family and together with it the evolution other kingdoms. However, this hypothesis still needs to be tested by assessing radiation patterns of the different Brassicaceae lineages, including *Aethionema*.

Our analyses place the crown node age of the Brassicaceae in the Early Eocene, around 48 Mya. This is congruent with the estimates of Beilstein *et al.* (2010; ~54.3Mya), Huang *et al.* (2015; ~42 Mya), Couvreur *et al.* (2010; ~38 Mya) and Cardinal-McTeague *et al.* (2016; ~43.4 Mya) and earlier than found by Edger *et al.* (2015) and Hohmann *et al.* (2015) (both ~ 32 Mya). There are two main methods to calibrate the age of divergence of a phylogeny, directly through fossils (primary calibration) or by using already published estimates (secondary calibration). When a secondary calibration method is used, the error rates should be taken into account (Franzke *et al.* 2016). The presently known Brassicaceae fossils are from derived branches in the core group (see Franzke *et al.* (2016) for an overview). As our focus here is on the divergence of *Aethionema*, a lineage without known fossils, we used the most recent calibration, including error rates, for the most recent common ancestor of the Brassicaceae and Cleomaceae (Cardinal-McTeague *et al.* 2016). Our estimates of the divergence of *Aethionema* into three clades corresponds to the estimations done until now (Hohmann *et al.* 2015, Huang *et al.* 2016, Cardinal-McTeague *et al.* 2016), taking the confidence intervals into account (Franzke *et al.* 2016).

We used a combination of silica dried and herbarium specimens to infer the chloroplast and nuclear rDNA phylogenies. The newest sequencing techniques make it possible to use natural history collections in a phylogenomic context, although fewer reads are recovered from older specimens (Bakker *et al.* 2015). Herbarium specimens have been used to

investigate the movement of domesticated crops (Ames & Spooner 2008) and their pests (Yoshida *et al.* 2013, Yoshida *et al.* 2014). Moreover, specimens in natural history collections can be from species that have been encountered only once (Sebastian *et al.*, 2010), are now extinct (Zedane *et al.* 2016) or occur in areas that are hard to access, due to geographical or political issues, as is the case here. However, as already known for a very long time (but quantified by Goodwin *et al.* (2015)) the uncritical use of natural history specimens can lead to inaccurate results. Our herbarium samples were curated under supervision of the authors (Al-Shehbaz, Dönmez) and 41% of the herbarium sheets used for the ancestral area reconstruction of *Aethionema* were identified by one of our researchers (Al-Shehbaz, Dönmez) in the last fifteen years. The latitude and longitude data used for the ancestral area reconstruction were critically assessed by hand from all the available digitized herbarium sheets to be certain of the correct representation in the data set (Suppl. Fig. 1).

To conclude, we found that the centre of origin of the genus *Aethionema* is most probably along the Anatolian Diagonal. *Aethionema* dispersed throughout the Irano-Turanian region after the uplift of the Iranian and Anatolian plateaus and the formation of the mountain ranges in Iran and Turkey. Being the sister lineage of the Brassicaceae core group, *Aethionema* comprise an evolutionary important position for comparative studies within the Brassicaceae. Understanding its phylogenomic history makes it possible for future researches to understand trait evolution within the Brassicaceae.

### Acknowledgements

We thank Dr. F.T. Bakker from Wageningen University and Research for the fruitful discussion during the study. Moreover we also want to thank the herbaria of Berlin, Edinburgh and Munich for providing samples for this study. This work was supported by the grants from NWO Vernieuwings Impuls VIDI (Grant number: 864.10.001) and (849.13.004), the latter as part of the ERA-CAPS “SeedAdapt” consortium project ([www.seedadapt.eu](http://www.seedadapt.eu)).

### Supplemental files

All Supplemental files are available upon request.

### Tables

Suppl. Table 1. Specimens used.

Suppl. Table 2. Table of sampled *Aethionema* specimens, their location or herbarium number and assembly descriptives and the plastome and rDNA genes used. *Aethionema spinosum*\* is a herbarium sample that our molecular data show as *A. spinosum* while its original sheet denotes *A. transhyrcanum*. Sheet 3 shows the coordinates used for the historical biogeographical analyses.

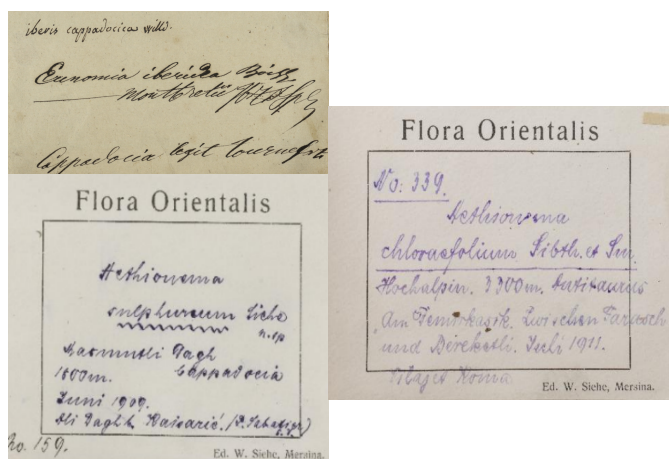
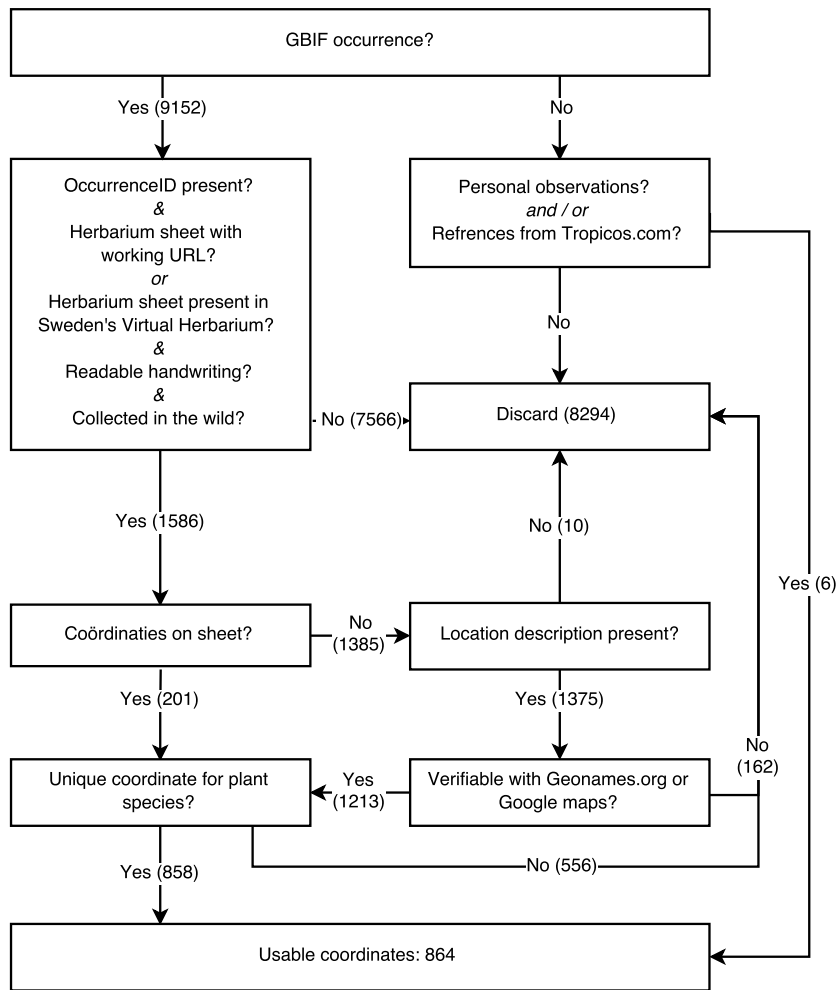
### Datasets

Suppl. Dataset 1. Concatenated alignment of the 76 chloroplast coding regions.

Suppl. Dataset 2. Concatenated alignment of nuclear rDNA coding and non-coding regions.

Suppl. Datasets 3 and 4. Original RASP results.

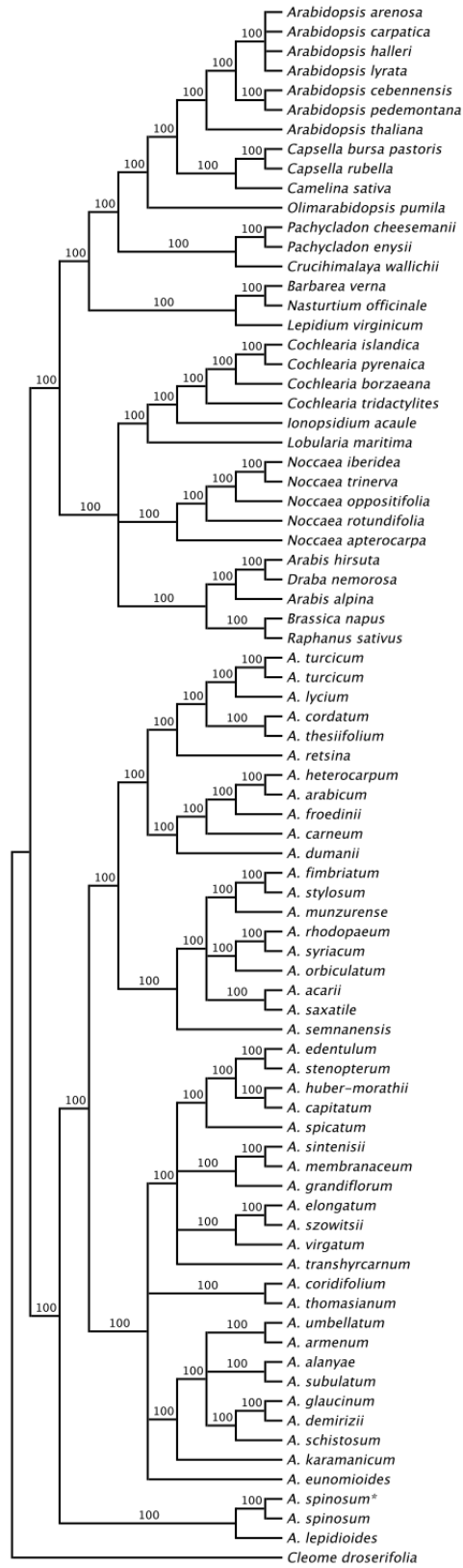
3

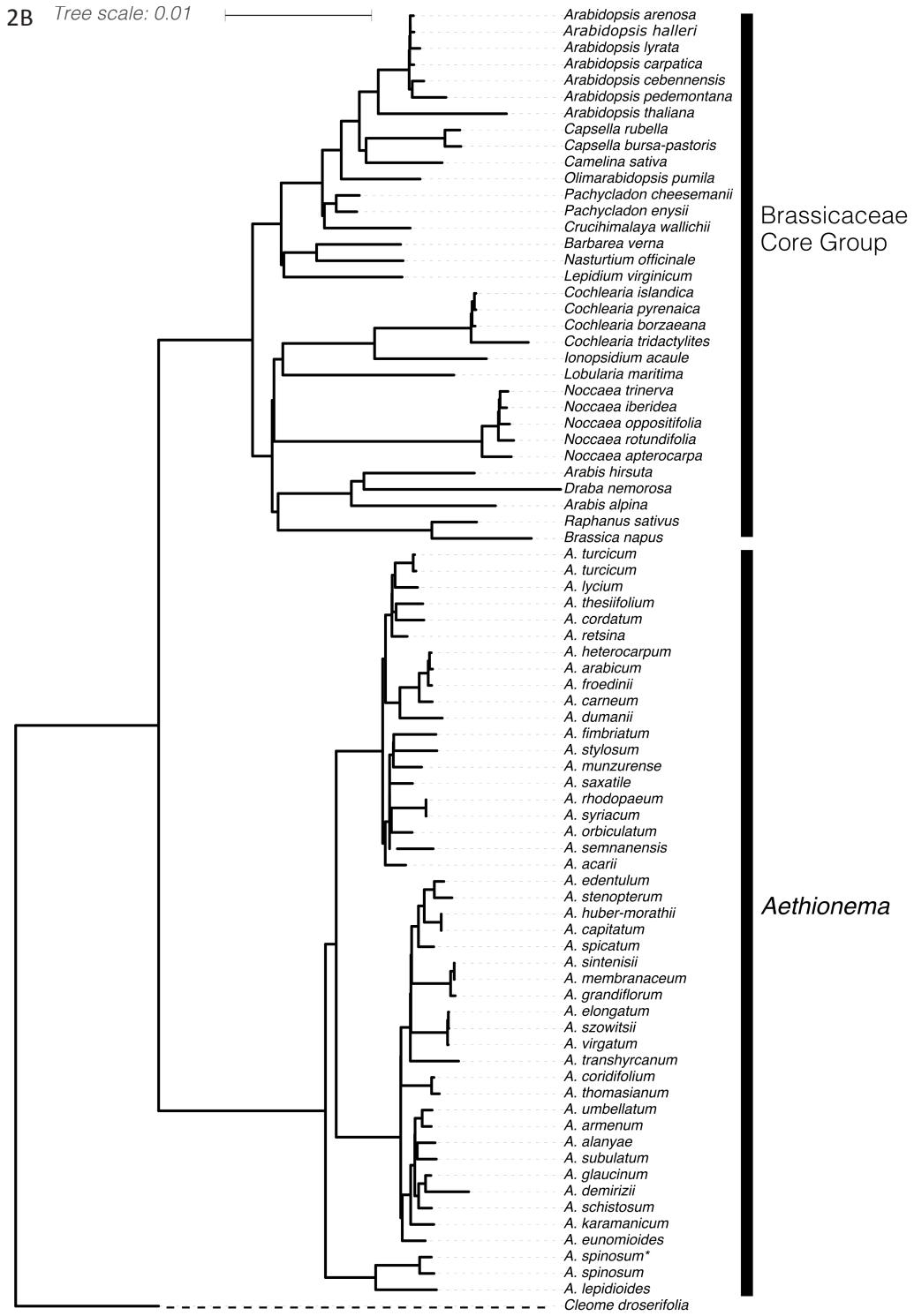


**Suppl. Fig. 1** Flow diagram for processing locality data obtained for historical biogeographical analyses. Numbers in brackets along the arrows show the amount of specimens that went through to the next step. Pictures below the diagram are examples of hand written and read herbarium sheet labels.



2A

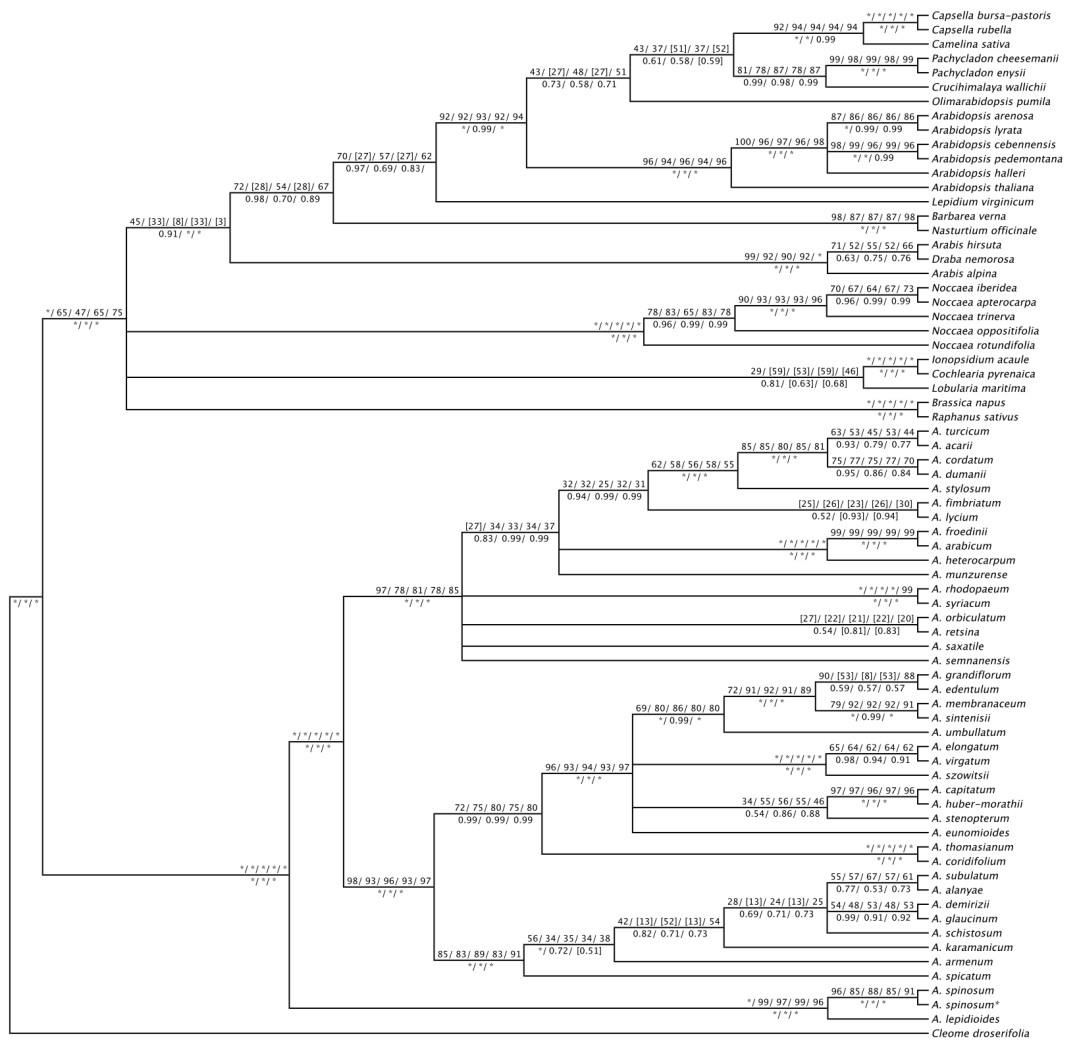




3

**Suppl. Fig. 2** A) Maximum Parsimony cladogram based on the chloro-matrix. B) Phylogram of the chloro-matrix (6 chloroplast coding regions) (maximum likelihood analyses, GTR+GAMMA, unpartitioned).

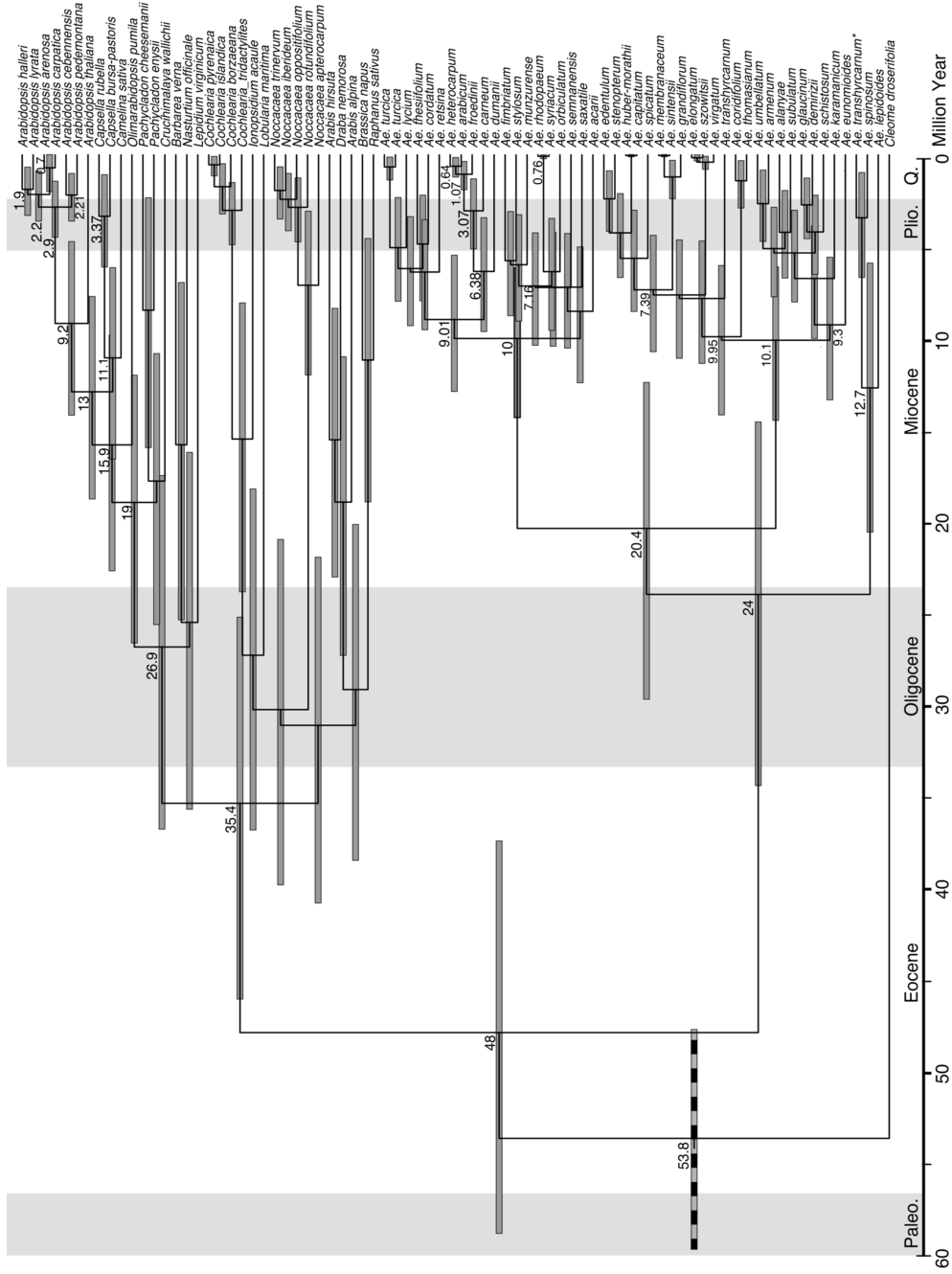
Phylogeny and historical biogeography



3

**Suppl. Fig. 3** Cladogram based on the maximum likelihood analysis (GTR+GAMMA, unpartitioned) of the rDNA-matrix. Bootstrap values for different partition models (unpartitioned/ITS vs coding region/ per gene partition/ PartitionFinder/ per codon partition) are shown above the branches. Posterior probabilities of the Bayesian inference based on different partition models (unpartitioned/ ITS vs Coding/ per gene partition) are shown below the branches. \* = 100% bootstrap support or a posterior probability of 1.

3



Suppl. Fig. 4 Chronogram of Bayesian dating analyses (BEAST). Horizontal bars represent 95% highest posterior density intervals around mean node ages. The hatched horizontal bar represents the secondary constrain on the root node height following Cardinal-McTeague *et al.* (2016).

## Chapter 4

### Genome-wide nucleotide diversity and associations with geography, ploidy level and glucosinolate profiles in *Aethionema arabicum* (Brassicaceae)

Setareh Mohammadin<sup>1</sup>, Wei Wang<sup>2</sup>, Ting Liu<sup>2</sup>, Hamid Moazzeni<sup>3</sup>,  
Kuddisi Ertugrul<sup>4</sup>, Tuna Uysal<sup>4</sup>, Charalambos S. Christodoulou<sup>5</sup>,  
Patrick P. Edger<sup>6</sup>, J. Chris Pires<sup>7</sup>, Stephen I. Wright<sup>2</sup>, M. Eric Schranz<sup>1</sup>

PLOS ONE, *under review*

---

<sup>1</sup> Biosystematics, Plant Science Group, Wageningen University and Research, Wageningen, The Netherlands

<sup>2</sup> Department of Ecology and Evolutionary Biology, University of Toronto, Toronto, ON, Canada

<sup>3</sup> Department of Botany, Research Center for Plant Sciences, Ferdowsi University of Mashhad, Mashhad, Iran

<sup>4</sup> Department of Biology, Faculty of Science, Selcuk University, Konya, Turkey

<sup>5</sup> Department of Forests, Ministry of Agriculture, Natural Resources and Environment, Nicosia, Cyprus.

<sup>6</sup> Department of Horticulture, Michigan State University, USA

<sup>7</sup> Division of Biological Sciences, University of Missouri, Columbia, USA

### **Abstract**

*Aethionema* species are the sister-group to the core Brassicaceae (including *Arabidopsis thaliana*) and thus have an important evolutionary position for comparative analyses. *Aethionema arabicum* (Brassicaceae) is emerging as a model to understand the evolution of various traits. We generated transcriptome data for seven *Ae. arabicum* genotypes across the species range including Cyprus, Iran and Turkey. Combined flow cytometry and single nucleotide polymorphism (SNP) analyses identified distinct tetraploid (Iranian) and diploid populations (Turkish/Cypriot). The Turkish/Cypriot lines had a higher genome-wide genetic diversity than the Iranian lines. However, one genomic region contained genes with a higher diversity in the Iranian than the Turkish/Cypriot lines. Fifteen percent of the genes in this region were chaperonins involved in protein folding. Moreover, a genome-wide selection analysis found evidence for positive selection of the paralogous *GSTF9/10* genes involved in responses to oxidative and drought stress. Additionally, an analysis of glucosinolate profiles, chemical defence compounds of the Brassicaceae, showed a difference in diversity of indolic glucosinolates between the Iranian and Turkish/Cypriot lines. Hence the different *Ae. arabicum* populations, Iranian vs. Turkish/Cypriot, both show potential adaptations to their local environments.

## Introduction

The genus *Aethionema* W.T. Aiton (tribe Aethionemeae) is the sister-group to the rest of the Brassicaceae family and thus serves as an important group to study the evolution of cruciferous traits. The fast-flowering annual species *Aethionema arabicum* (L.) Andr. ex DC. (Aethionemeae, Brassicaceae) is being utilized as a model for comparative analysis. For example, genetic mapping of *Ae. arabicum* provides the opportunity to understand structural genomic evolution in the light of Brassicaceae genomic blocks (Lysak *et al.* 2016, Nguyen *et al.* in prep). Moreover, *Ae. arabicum* is heterocarpic, and completes its life cycle between April and the end of June on the steep stony slopes of the Irano-Turanian region (Lenser *et al.* 2016, Bibalani 2012). Heterocarpy is defined as the occurrence of two types of fruits on the same infructescence whereby the fruits and seeds differ in size, colour, shape, dormancy and germination (Lenser *et al.* 2016, Imbert 2002). The short life-cycle and the heterocarpic phenotype are likely adaptations to the unpredictable local growth conditions of *Ae. arabicum*.

*Aethionema arabicum* is native to the stony slopes across the Irano-Turanian region with its most Western and Northern border being Bulgaria (Velchev 2015). The Irano-Turanian region harbours three biodiversity hotspots and is at the interface of the Mediterranean, Saharo-Sindian and Euro-Siberian regions containing 32,000 endemic species (Takhtajan 1986, Manafzadeh *et al.* 2016, Davis 1965). In the western Irano-Turanian region a floristic break occurs that connects the south-western Turkish mountains with the north eastern Iranian mountains: the Anatolian Diagonal (Davis 1965). Hence the mountains of the Anatolian Diagonal connect the otherwise isolated mainly west-eastern oriented Anatolian mountain ranges (Ansell *et al.* 2011). The Anatolian Diagonal has been regarded as a barrier for animals (Gül 2013, Vamberger *et al.* 2013) and a migratory route for plant species (Ansell *et al.* 2011, Mummenhoff *et al.* 2001, Jordon-Thaden 2009, Karl & Koch 2013). Twenty-four percent of the Brassicaceae species occur in the Irano-Turanian region (Koch & Kiefer 2006) and it has been hypothesized to be the “ancestral area” or centre of origin for the entire Brassicaceae (Al-Shehbaz *et al.* 2006, Couvreur *et al.* 2010, Franzke *et al.* 2011, Warwick *et al.* 2010, Hedge 1976). This hypothesis holds for the crucifer genera *Draba*, *Lepidium*, *Ricotia* and *Aethionema* (Mummenhoff *et al.* 2001, Jordon-Thaden 2009, Jordon-Thaden *et al.* 2013, Özüdoğru *et al.* 2015, Mohammadin *et al.* submitted).

The genus *Aethionema* shares the typical cruciferous traits, e.g. the methionine-derived glucosinolates (GS, i.e. mustard oils) with the Brassicaceae core group (Edger *et al.* 2015, Hofberger *et al.* 2013). GS are a novel suite of metabolites developed within the Brassicales as chemical defences against pathogens and herbivores (Hofberger *et al.* 2013, Edger *et al.* 2015, Halkier & Gershenzon 2006). Glucosinolates are derived from a basic sugar molecule and an amino acid with a side chain that can be elongated with carbon molecules (Halkier & Gershenzon 2006, Redovnikovic *et al.* 2008). In addition to their function as defence compounds, GS can also act as attractants and are economically used for their mustard flavour and anti-carcinogenic activity (Halkier & Gershenzon 2006). Although all Brassicales contain GS, the At-alpha WGD event and the arms race between Brassicaceae and its Pieridae herbivores likely increased the GS diversity of Brassicaceae (containing 120+ different compounds (Edger *et al.* 2015)).

In addition to the phenotypic synapomorphies shared with the rest of the Brassicaceae, *Aethionema* also share the At-alpha WGD event with the rest of the Brassicaceae (Schranz *et al.* 2012, Edger *et al.* 2015). WGD, also known as polyploidy, can also be associated with shifts in speciation rates (Koenig & Weigel 2015, Soltis *et al.* 2009, Zhang 2003, Ha *et al.* 2009). Genes duplicated by WGD can go through the process of pseudogenization (gene loss), subfunctionalization (partitioning of ancestral functions), neofunctionalization (novel gene function), and/or retain their original gene function (Song *et al.* 1995, Zhang 2003). Whatever their future may be, duplicated genes provide new material for mutations and therefore new material for natural selection and genetic drift to act upon (Koenig & Weigel 2015).

Polyploidisation can occur by a lack of gamete reduction (autopolyploidy) or by genome hybridization of two species coupled with lack of gamete reduction (allopolyploidy). After an allopolyploid event, polyploid and diploid progenitors have differing chromosome complements, leading to potential immediate reproductive isolation. However reproductive isolation is not always the case. For example, gene flow from the diploid *Capsella rubella* (Brassicaceae) to the tetraploid *Capsella bursa-pastoris* contributes to the genetic variation of *C. bursa-pastoris* (Slotte *et al.* 2008). Although a WGD event could be a playground for natural selection, many studies show that polyploids are not significantly differentiated from their diploid relatives (Arrigo & Barker 2012), nor are their speciation rates higher (Mayrose *et al.* 2011).

## 4

*Aethionema arabicum* has mainly been studied in a comparative framework with the rest of the Brassicaceae (Mohammadin *et al.* 2015, Hofberger *et al.* 2013, Beilstein *et al.* 2012), to investigate the genome and transcriptome sequences (Haudry *et al.* 2013, Edger *et al.* 2015, Mohammadin *et al.* 2015), long non-coding RNAs and other conserved non-coding sequences (Haudry *et al.* 2013, Mohammadin *et al.* 2015), telomerases (Beilstein *et al.* 2012), and the glucosinolate biosynthetic pathway (Hofberger *et al.* 2013). To improve our understanding of *Ae. arabicum* evolution, knowledge about the patterns of genetic diversity and structure within the species are required. Here we used a genome-wide approach to understand the genomic diversity between different *Ae. arabicum* accessions sampled widely in the Irano-Turanian region. Analysis of nucleotide polymorphism data derived from *de novo* transcriptome sequencing reveal two clear geographical groups that correlate with ploidy differences found by flow cytometric analyses. Furthermore, GS analysis finds that the geographical clusters differ in their defensive GS profiles. Finally, we also find differences in selection patterns of particular genes across the genome.

## Materials and Methods

### Plant material, RNA isolation, sequencing and assembly

*Aethionema arabicum* seeds from seven different accessions from Cyprus, Iran and Turkey were sown in sowing-soil in pots (9cmx9cmx10cm) and grown in the greenhouse at the University of Amsterdam (18°C at night, 20°C day temperature, 12:12 light:dark regime) in the winter of 2011. Turk1 and Turk2 were from Konya (Turkey, UTM coordinates: 36.58077 N; 032.27649 E and 37.01166 N; 032.19826 E), Turk3 from Elaziğ, Turkey (no GPS data available). Iran1 came from the Dizin mountains (Karaj, Iran, UTM coordinates: 36.06851 N; 051.19645 E) and Iran2 and Iran3 from the base of the Touchal mountain (Tehran, Iran, UTM



coordinates: 36.06851 N; 051.19645 E). Cyp seeds were from Kato-Moni (Cyp, Cyprus, UTM coordinates 35.326335 N; 33.530596 E). Sampling localities are shown on the map of Fig. 1. RNA isolation, cDNA synthesis, library preparation and assembly were done on the same tissues and with the same procedure as for *Ae. arabicum* in (Mohammadin *et al.* 2015). The lines were single-end sequenced with the Illumina Hiseq2000 sequencer on 1x100bp lanes, with 3 lines per lane. This resulted in 5Gbp of reads per line. *Aethionema carneum* (Banks & Sol.) B. Fedtsch, also a heterocarpic annual belonging to the same clade as *Ae. arabicum*, was used as the out-group (Lenser *et al.* 2016) for selection pattern analyses. The *Ae. carneum* transcriptomes generated by Mohammadin *et al.* (2015) were used.

### Flow cytometry

In addition to the seven *Ae. arabicum* plants mentioned above, we sowed *Ae. carneum* and the *Ae. arabicum* reference line for flow cytometry measurements (see above for sowing details). The ploidy levels of the eight *Ae. arabicum* lines and *Ae. carneum* were measured with a flow cytometer (Partec Clab<sup>TM</sup>, Munster, Germany). Fully-grown leaves of the same age were used (gain = 340). The nuclei were isolated by mincing the leaves with a razor blade in 1ml of the standard isolation buffer 4,6'-diamidino-2-phenylindole (DAPI) (Dolezelt *et al.* 1998).

**Table 1** Ploidy level and percentage of heterozygosity of *Aethionema arabicum* lines.

Line	Ploidy (2x)	Heterozygosity (%)
<i>Ae. carneum</i>	2	1.36
Cyp	2	1.61
Turk1	2	1.47
Turk2	2	0.83
Turk3	2	1.24
Iran1	4	2.20
Iran2	4	2.76
Iran3	4	2.70

### SNP calling and filtering

Single Nucleotide Polymorphisms (SNPs) were called with the unified genotyper of GATK (McKenna *et al.* 2010, DePristo *et al.* 2011) with the BadCigar read filter, operating over scatter\_005 intervals and using the *Ae. arabicum* genome (Haudry *et al.* 2013) as a reference. Alignments of the coding regions were made using in house perl scripts producing diploid alignments. The VCF file was filtered with GQ  $\geq$  40, excluding SNPs with missing alleles. VCFtools (Danecek *et al.* 2011) was used to filter SNP data, and calculate the number of heterozygous loci per individual and fixed heterozygosity levels. Diploid alignments were made based on the GATK output, with 'allele\_1' being the same as the reference and 'allele\_2' containing all the mutations.

To further validate the assignment of potential alleles for heterozygous regions, we additionally called the alleles based on phased SNPs with HapCut (Bansal & Bafna 2008). SNPs with a GQ  $\leq$  60 were removed before the following analysis. All analyses were done using in-

house perl scripts and MySQL. Phased haplotype blocks were based on the HapCut output (heterozygous SNPs) and the VCF output (homozygous SNPs). Since some of our populations are tetraploid (see Results), we wished to investigate the possibility of allopolyploid origins. However, without knowing the diploid parental genomes, it is not feasible to phase across distinct haplotype blocks. Hence, to investigate the possibility of allopolyploidy we assumed that the haplotypes with more alleles from the reference genome are from one chromosome sequence (allele\_1) and the haplotypes with more alternative alleles are from the second chromosome (allele\_2).

### Population structure analyses

STRUCTURE v 2.3.4 (K=2- 5, burn-in period=10,000, 20,000 runs; Pritchard *et al.* 2000) and PCOA analyses were used for population structure analyses. Both were conducted on a genotype matrix generated with VCFtools ('-O12' command) on the filtered SNPs. The reference line was added to this matrix as an individual with only homozygous reference alleles. The PCOA analysis was done with the "bigmemory" and "ape" packages in R v 3.2.1 (Kane *et al.* 2013, Paradis *et al.* 2004). With the exception of the network tree all figures were made in R v 3.2.1 with the "ggplot2" package (Wickham 2009) and combined with inkscape and GIMP.

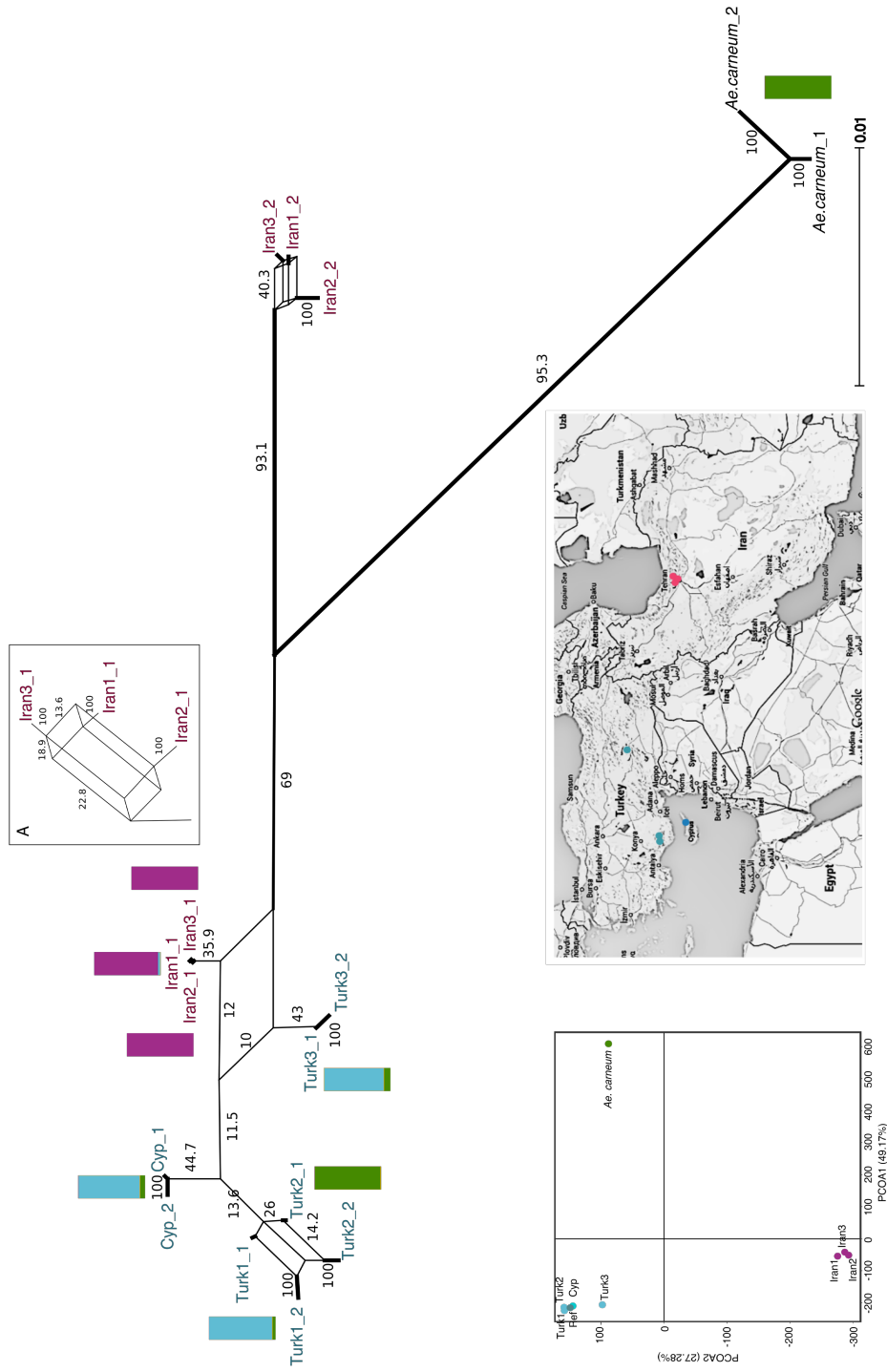
Maximum likelihood trees were inferred for the unphased alignments that passed the SNP filters (RAxML v8.2.4, GTR-GAMMA nucleotide model, 1000 rapid bootstraps, random seed=12345 and both *Ae. carneum* alleles as outgroups, (Stamatakis 2014)). A consensus network was made from all the maximum likelihood trees in SplitsTree (mean edge-weight and threshold = 0.1, Huson & Bryant 2006).

To assess whether the network pattern does or does not depend on tandem duplicates, single copy genes of *Ae. arabicum* were used for a separate consensus network analyses. We assessed synteny between *Ae. arabicum* and the 959 *Arabidopsis thaliana* single copy genes (Duarte *et al.* 2010) with SynFind in CoGe (Lyons & Freeling 2008). Genes from the unphased alignments with only one syntenic hit and no proxys that occurred in our gene set were used.

We used the first twenty-five genes of the phased dataset to assess whether their network differs from the unphased alignments. However, as these genes gave us the same result as the unphased network we continued with the unphased dataset.

### Genome-wide diversity analyses

Following the results of STRUCTURE, PCOA and gene tree network analyses, we further analysed the unphased data. We analysed each allele type separately and split the individuals in two groups: the Iranian group and the Turkish group (including the Cypriotic line) based on ploidy differences. The diploid alignments were separated into allele type with an in-house python script. Population genetic summary statistics were obtained with Phylomorphorama (Bachtrog & Andolfatto 2006) for every allele type separately and averaged per allele to have a per gene per individual specific descriptive. For genome-wide patterns we used our



**Fig. 1** Consensus network of 8,969 bi-allelic expressed protein-coding genes of *Aethionema arabicum* lines with *Aethionema carneum* used as the outgroup. The numbers along the branches show the percentage of trees supporting the branch. Naming of the tips is as follows: LocationPopulation\_Allele. Hence Turk1\_1 and Turk1\_2 are two alleles of the Turk1 population, coming from Turkey. The inserted map shows the population clusters with the percentage of variation by the axis within brackets, the red one to the Iranian individuals. Inserted is a PCOA plot showing the population clusters with the percentage of variation by the axis within brackets. The bars along the network show the result of a STRUCTURE analysis for K=3, only one bar per population is shown. (A) Shows the structure of the IranX\_1 allele.

**Table 2** Population genomic descriptives for Turkish and Iranian lines of *Aethionema arabicum*. *Aethionema carneum* was used as the outgroup to calculate the values for segregating sites. N= Total number of the site class, S = number of polymorphic sites, D= number of segregating sites. Averages are shown for the synonymous nucleotide diversity ( $\pi$ ) and the pairwise divergence (Dxy) with SD in brackets.

Group	Site Class	N <sub>sites</sub>	S	D	$\pi$	D <sub>xy</sub>
Turkey (n=4)	Synonymous	2,545,326.617	46737.5	130091.5	0.0108 (0.0156)	0.0636 (0.0338)
	Nonsynonymous	8,535,539.384	20787.5	51385	0.0014 (0.0025)	0.0075 (0.0074)
Iran (n=3)	Synonymous	2,624,706.333	4141	171352	0.0012 (0.0039)	0.0705 (0.0301)
	Nonsynonymous	8,774,195.667	4060	70775.5	0.0003 (0.0012)	0.0087 (0.0076)

genetic linkage map (Nguyen *et al.*, in prep). dN/dS between *Ae. arabicum* and *Ae. carneum* was directly calculated from the Phylomorphorama output using the ratio of per site non-synonymous to synonymous substitution rates, also first per allele type and then averaged over both types to get a per gene dN/dS value. Sliding windows were calculated for windows of 50cM with a step size of 8cM using an in-house R script and the genetic map of Nguyen *et al.* (in prep). *Arabidopsis thaliana* homologs of the *Ae. arabicum* genes were assessed with the SynFind tool CoGe (Lyons & Freeling 2008).

## 4

### Glucosinolate isolation and analysis

Seeds from all lines were imbibed on wet filter paper in the dark for three days at 20°C. Five seeds from the same individual were sown in 12cm round pots, in sowing soil, in the greenhouse at Wageningen University and Research in the summer of 2016. Flowers, fruits and leaves were isolated from 3 adult plants per line, frozen in liquid nitrogen and kept at -80°C. Samples were freeze-dried for 24 hours and ground with 4-10 2mm glass beads. Samples that were lighter than 4mg were pooled. Samples between 4-10mg were extracted with 1mL of 80% methanol with an internal standard of 0.05 mM intact 4-hydroxybenzylglucosinolate. The samples were analysed by HPLC-UV following Burow *et al.* (2006) with the following adjustments of the chromatographic gradient: water (A)-acetonitrile (B) gradient (0-8 min, 10-50% B; 8-8.1 min, 50-100% B; 8.1-10min 100% B and 10.1-13.5 min 10% B; flow 1.0 mL min<sup>-1</sup>). The *Ae. arabicum* genes of the glucosinolate pathway were based on the *A. thaliana* homologs found by (Hofberger *et al.* 2013). Using BlastN we assessed whether these genes were expressed in our transcriptome data and whether they were under selection by assessing their presence in our list of genes with SNPs.

## Results

### Sequencing statistics and flow cytometry

We sequenced the transcriptomes of seven *Ae. arabicum* lines that are representative of the species distribution range. In addition, we used the already published transcriptome of the reference line of *Ae. arabicum* (Haudry *et al.* 2013) and of the outgroup species *Ae. carneum* (Mohammadin *et al.* 2015).

The *Ae. arabicum* accessions from Turkey, Cyprus, Iran and *Ae. carneum* had an average of 38,689,493.8 ( $\pm 7,951,220.4$  SD) reads before quality trimming. 87.3% of the reads remained after quality trimming (33,776,390.2  $\pm$  6,958,395.1 SD). These were assembled relative to the reference line (Haudry *et al.* 2013) into an average of 18,881 ( $\pm$  210.6 SD) contigs with an average length of 803 bp ( $\pm$  3.1 SD).

The heterozygosity levels of the Iranian lines exceeded the percentage of heterozygous loci between the *Ae. arabicum* reference line and the outgroup *Ae. carneum* (Table 1). Flow cytometry analyses showed that the Iranian lines are tetraploid while all the other *Ae. arabicum* lines and *Ae. carneum* are diploid, (Table 1). Thus, the higher degree of observed 'heterozygosity' of the Iran lines is caused by (disomic) polyploidy.

### Population structure

A total of 171,916,211 SNPs were called for all individuals against the reference *Ae. arabicum* genome. Filtering reduced this to a total of 22,088,876 bi-allelic sites, representing 9,070 coding genes.

The PCOA and STRUCTURE analyses show the clustering of the Iranian and the Turkish lines (Suppl. Fig. 1, eigenvalues PCOA1=610,115.6, PCOA2=346,738.7, explaining a total of 76.45% of the variation). The PCOA results also show that the reference line is nested with other Turkish lines (Fig. 1).

Out of the 9,070 coding genes with SNPs, there were 8,969 unphased alignments that passed the RAXML criteria. The consensus network of all 8,969 trees show that in 93.5% of the trees the second allele from the Iranian lines split off from the other alleles (Fig. 1). Both alleles of the Turkish lines, the first allele of the Iranian lines and the Cypriot alleles cluster together. Only 11.5% of the trees support a separation of the first Iranian allele from the Turkish and Cypriot alleles (Fig. 1). This indicates a similarity of one of the Iranian alleles with the primary Turkish/Cypriot allele cluster, while the other allele is as separated from the Turkish/Cypriot cluster similar to the outgroup species *Ae. carneum*. This separation suggests a likely tetraploid origin or an introgression history for the Iranian lines. The network made of 158 single copy genes and of the phased genes showed the same pattern (Suppl. Fig. 2, Suppl. Fig. 3).

Following the above-mentioned ploidy differences between the Iranian and Turkish/Cypriot (hence called Turkish) lines, we analysed the two geographical groups separately for their genome wide diversity patterns. This was conducted to exclude biases by quantitatively comparing populations with different ploidy levels, although this also assumes a potential allopolyploid origin of the Iranian lines by analysing diversity for the Iranian lines within putative allelic homeologs.

### Population genomic analyses

The dataset of the Turkish individuals contained a total of 11,080,866 sites, of which 0.61% was polymorphic and 1.64% diverged from *Ae. carneum* (Table 2). The Iranian individuals had a total 11,398,902 sites, with 0.07% polymorphic sites and 2.14% sites diverging from

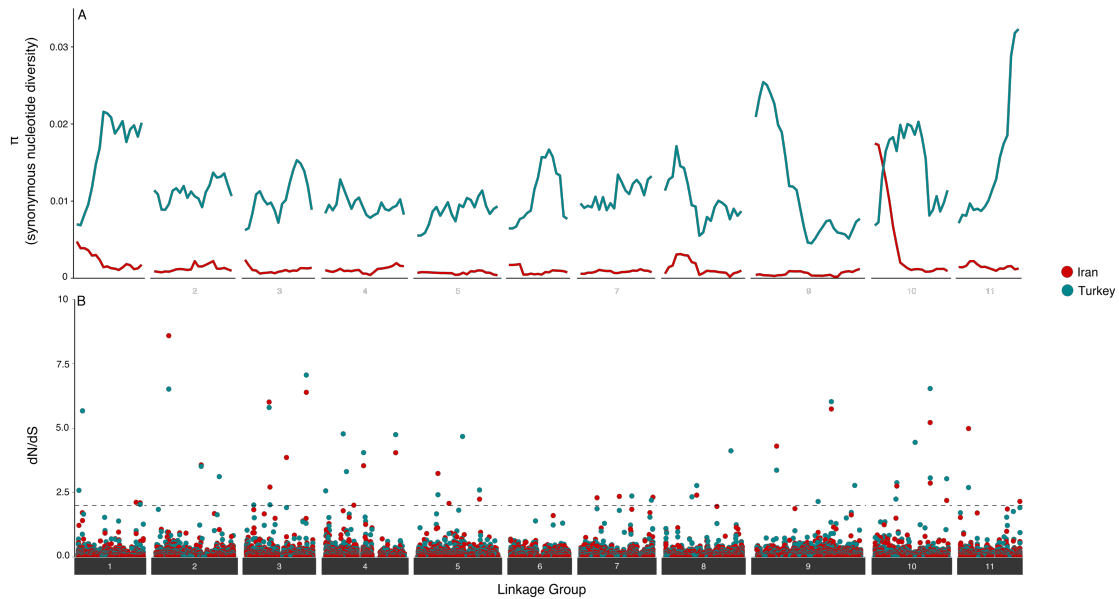
Chapter 4

**Table 3** Forty-two *Aethionema arabicum* (Gene) genes with the dN/dS values above 2 and the function of their *Arabidopsis thaliana* (Ath) homologs. Ordering is based on linkage group (LG) of Nguyen *et al.* (in prep.)

Gene	d N / d S Iran	d N / d S Turkey	LG	Ath homolog	Function
AA46G00131	2.12	NA	1	AT2G22420	Peroxidase superfamily protein
AA46G00063	2.1	2.04	1	AT2G21390	Coatomer, alpha subunit
AA39G00033	NA	5.67	1	AT1G17340	SAC5; Phosphoinositide phosphatase family protein
AA37G00130	NA	2.59	1	AT1G16250	Galactose oxidase/kelch repeat superfamily protein
AA38G00092	3.58	3.52	2	AT1G69890	Unknown function
AA31G00578	8.6	6.5	2	AT1G73660	SIS8; similarity to MAPKKs. May function as a negative regulator of salt tolerance.
AA16G00008	NA	3.12	2	AT3G46960	SKI2; The gene encodes a DEXD/H box RNA helicase
AA33G00186	6.4	7.06	3	AT2G18915	ADAGIO2; encodes a member of F-box proteins
AA26G00586	2.71	2.03	3	AT3G19130	RBP47B; RNA-binding protein 47B
AA26G00557	6.01	5.8	3	AT3G1878	ACT2; an actin constitutively expressed in vegetative structures
AA26G00211	NA	2.02	3	AT3G14990	ATDJA1A; a homolog of animal DJ-1 superfamily protein
AA17G00105	3.86	NA	3	AT3G23080	Polyketide cyclase/dehydrase and lipid transport superfamily
AA60G00134	4.05	4.75	4	AT2G36580	Pyruvate kinase family protein
AA32G00948	2.01	NA	4	AT2G27180	Unknown protein
AA32G00681	NA	3.32	4	AT2G30170	PCBP; a chloroplast PP2C phosphatase required for efficient dephosphorylation of PSII proteins, involved in light acclimation
AA32G00635	NA	4.78	4	AT2G30860	GSTF9; Encodes glutathione transferase from phi class of GSTs
AA32G00024	NA	2.57	4	AT4G00752	UBX domain-containing protein
AA10G00445	3.54	4.05	4	AT4G26850	GGP; involved in ascorbate biosynthesis
AA53G00888	3.25	2.42	5	AT3G26782	Tetratricopeptide repeat (TPR)-like superfamily protein
AA53G00570	2.08	NA	5	AT1G61740	Sulfite exporter TauE/SafE family protein
AA53G00165	NA	4.68	5	AT1G32080	Encodes LrgAB/CidAB protein localized to the chloroplast envelope involved in chloroplast development, carbon partitioning and leaf senescence.
AA102G00285	2.25	2.6	5	AT4G31340	Myosin heavy chain-related
AA87G00277	2.3	NA	7	AT4G14410	BHLH104; DNA binding
AA6G00107	2.35	NA	7	AT4G19860	Encodes a cytosolic calcium-independent phospholipase A.
AA3G00143	NA	2.37	7	AT4G09970	Unknown protein

Table 3 continued

Gene	d N / d S Iran	d N / d S Turkey	LG	Ath homolog	Function
AA27G00216	2.32	NA	7	AT5G35910	RPP6L2; a nuclear-localized RRP6-like protein
AA27G00041	NA	2.21	7	AT5G38160	Bifunctional inhibitor/lipid-transfer protein/seed storage 2S albumin superfamily protein
AA55G00047	2.4	2.77	8	AT5G52060	A member of Arabidopsis BAG (Bcl-2-associated athanogene) proteins.
AA44G00394	NA	4.13	8	AT5G11420	Unknown protein
AA41G00046	2.34	2.34	8	AT5G51150	Mitochondrial import inner membrane translocase subunit
AA65G00058	4.3	3.37	9	AT1G09970	RLK7 belongs to a leucine-rich repeat class of receptor-like kinase (LRR-RLKs). Involved in the germination speed and the tolerance to oxidant stress.
AA4G00198	NA	2.15	9	AT5G64650	Ribosomal protein L17 family protein
AA21G00354	NA	2.78	9	AT2G46200	Unkown protein.
AA19G00252	5.75	6.03	9	AT5G03240	UBQ3; encodes ubiquitin that is attached to proteins destined for degradation.
AA62G00082	NA	4.45	10	AT3G53360	Tetratricopeptide repeat (TPR)-like superfamily protein
AA61G00625	2.2	3.04	10	AT3G62980	Tetratricopeptide repeat (TPR)-like superfamily protein
AA61G00276	2.86	3.07	10	AT3G58570	P-loop containing nucleoside triphosphate hydrolases superfamily protein
AA61G00272	5.22	6.54	10	AT3G58510	DEA(D/H)-box RNA helicase family protein
AA2G00087	NA	2.25	10	AT1G49540	ELP2; elongator protein 2
AA2G00072	2.75	2.88	10	AT1G49340	PI4K; phosphatidylinositol 4-kinase expressed in inflorescences and shoots
AA20G00085	2.16	NA	11	AT2G20920	Protein of unkown function
AA57G00218	4.99	2.69	11	AT4G19600	Encodes a cyclin T partner CYCT1;4. plays important roles in infection with Cauliflower mosaic virus (CaMV).



**Fig. 2** Synonymous nucleotide diversity and the dN/dS ratio along the linkage map (Nguyen et al 2015) of Iranian (red) and Turkish/ Cypriot (blue) *Aethionema arabicum* populations. A) Iranian (red) and Turkish (blue) nucleotide diversity of protein coding genes over a sliding window over 50cM with a step size of 8cM. B) dN/dS ratios of Iranian (red) and Turkish (blue) protein coding genes. Every dot represents a gene.

## 4

*Ae. carneum* (Table 2). The synonymous polymorphism of the Iranian lines was much lower than the Turkish lines ( $\pi_{\text{Iran}} \sim 0.12\%$  vs.  $\pi_{\text{Turk}} \sim 1.08\%$ , Table 2), although the Iranian lines come from a very restricted geographical region. The levels of polymorphism of non-synonymous sites were decreased by an order of magnitude for both groups ( $\pi_{\text{Iran}} \sim 0.03\%$  vs.  $\pi_{\text{Turk}} \sim 0.14\%$ , table2).

### Genome-wide selection patterns

The within population polymorphisms, synonymous substitution rates, and the between species dN/dS ratios were calculated for every gene and plotted relative to their genomic location based on their genetic linkage map positions (Fig. 2). dN/dS ratios were calculated between the outgroup species *Ae. carneum* and the two populations (Iran and Turkey). Of the 9,070 genes that passed the SNP calling restrictions, only 6,867 could be mapped relative to the genome. Of the 2,203 genes that did not map against the linkage map, 587 genes were on scaffolds that have not yet been incorporated in the linkage map. The other 1,616 genes had coordinates beyond than the outer limits of the linkage groups and were hence excluded from additional analyses.

Synonymous nucleotide diversity measures are lower in the Iranian lines compared to the Turkish lines (see above and Table 2). However, at the beginning of linkage group 10 there is a peak in diversity of the Iranian lines that decreases within 10.6cM (Fig. 2A). The synonymous polymorphism level is significantly higher here than in the rest of the transcripts (average  $\pi_{\text{syn}} = 0.016 \pm 0.0096$  SD, Wilcoxon Rank Sum test with continuity correction,  $W=20482$ ,  $p < 2.2e-16$ ). Within the same 10.55cM interval the Turkish lines have lower but increasing



**Table 4** Glucosinolate content of Iranian, Turkish and Cypriot lines of *Aethionema arabicum*. Given are the GS concentrations in  $\mu\text{mol/g}$  with SD in brackets (n=3). Samples without a SD had to be pooled for GS analysis.

	Iranian			Turkish			Cypriot		
	Leaves	Flowers	Fruits	Leaves	Flowers	Fruits	Leaves	Flowers	Fruits
Aliphatic	3MTP	1.7 (1.1)	4.8	1.1 (0.7)	5.6	3.1 (4.1)	0.9	5.8	12.0
	3MSOP	2.7 (1.3)	0	1.6 (1.0)	3.2	2.8 (3.3)	0.7	0	12.9
	3MSOOP	0.6 (0.2)	0.3	2.0 (1.5)	0.4	1.1 (0.8)	1.1	0.4	2.0
	7MSOH	0 (0)	0	0 (0)	0.1	0.1 (0.1)	0.04	0.1	0.2
	8MSOO	0.4 (0.1)	1.1	4.1 (2.1)	1.7	6.4 (2.1)	1.03	2.1	6.0
Total	5.4	6.2	8.8	1.5	11	13.5	3.77	8.4	33.1
Indolic	I3M	0.1 (0)	0.1	0.1 (0)	0.1	0.3 (0.5)	0.08	0.1	0.2
	4MOI3M	0 (0)	0.1	0.1 (0)	0.1	0.3 (0.2)	0.03	0.1	0.4
	1MOI3M	0 (0)	0	0	0.1 (0.1)	0.1 (0.1)	0.004	0	0.4
	4OH13M	0 (0)	0	0	0 (0)	0	0.002	0	0
	Total	0.1	0.2	0.2	0.4	0.3	0.8	0.116	0.2

synonymous substitution rate, although this does not differ from the rest of the transcripts (average  $\pi_{\text{syn}} = 0.0074 \pm 0.0093$  SD, Wilcoxon Rank Sum test with continuity correction,  $W=216,580$ ,  $p=0.29$ ). This region harbours 59 genes with dN/dS averages of 0.1434 ( $\pm 0.156$  SD,  $n=57$ ) for the Iranian lines and 0.1434 ( $\pm 0.1666$  SD,  $n=55$ ) for the Turkish lines (Suppl. Table 1). dN/dS values were not significantly different from the rest of the genome (Wilcoxon rank sum tests with continuity correction, Iran:  $W=122,090$ ,  $p=0.945$ ; Turkey  $W=117.760$ ,  $p=0.6829$ ). However as the dN/dS values here are below one there is a signal of purifying selection. Most of the genes belong to an enzymatic group, are involved in DNA and/or RNA binding, zinc ion binding or transcription factors (Suppl. Table 1). There are also genes involved in seed maturation, embryogenesis, oxygen sensing, mitochondrion inner membrane proteins and chloroplast chaperonins (Suppl. Table 1).

The averages of non-synonymous versus synonymous substitutions, dN/dS, were significantly higher for Turkey 0.155 ( $\pm 0.322$  SD,  $n=6,691$ ) than for Iran: 0.1507 ( $\pm 0.287$  SD,  $n=6,774$ , Wilcoxon Rank Sum test with continuity correction,  $W=23,458,000$ ,  $p=4.153e-04$ ). Except for linkage group six, all the linkage groups had some genes that were under selection with a dN/dS above 2 (Fig. 2). There were a total 42 genes with evidence for being under strong positive selection, with dN/dS values above 2 (Table 3). Six out of these 42 genes are protein-coding genes within the organelles, mitochondrion and/or chloroplasts. Nineteen out of 42 genes are selected in both Iranian and Turkish lines. These genes encode fundamental cellular functions, e.g. F-box proteins and RNA helicase family proteins, but also genes that might be involved in responses to salt stress (Table 3). In the Turkish lines there is a strong selection for *GST9*, which is the tandem duplicate of *GSTF10*. These genes are involved in the building of the core structure of indolic glucosinolates and induced by drought stress (Sønderby *et al.* 2010, Ryu *et al.* 2009). Some of the genes under selection here have extremely high dN/dS values (Table 3), up to dN/dS > 8, these extreme values maybe due to misalignment between paralogous tandem duplicated genes.

### Differences in glucosinolate content

The Iranian and Turkish lines were very similar in their aliphatic glucosinolate content within their different tissues (Table 4). However, the indolic GS profiles varied: the Turkish and Cypriot lines have a higher diversity of GS components compared with the Iranian lines. Out of the 89 genes of the GS pathway that are syntenic with *Ae. arabicum* (Hofberger *et al.* 2013) 39 were expressed in our transcriptomes (Suppl. Table 2). Our filtered unphased SNP-list contained 17 *Ae. arabicum* syntelogs of the GS pathway (Suppl. Table 2), including genes from the entire biosynthetic pathway, ranging from the sulphatases SOT18, that are involved in the final step of GS compound formation to the side-chain changing GOX5 (Suppl Table 2) (Sønderby *et al.* 2010). These genes showed similar patterns of variation to the genome-wide analyses, with low genetic diversity for the Iranian lines and a higher synonymous  $\pi$  for the Turkish lines. However, the *Arabidopsis thaliana* homologs involved in the biosynthesis of the core structure of indolic GS, namely *GSTF9* and *GSTF10*, were under severe positive selection within the Turkish lines. *GSTF9/10* had a dN/dS = 4.78 within the Turkish lines and 1.79 within the Iranian lines.

## Discussion

Here we present the population structure, genetic diversity and selection pattern of *Ae. arabicum* using a genome wide approach and transcriptome analysis from a wide geographical range. Despite our small sample size, *Aethionema arabicum* populations have different ploidy levels with individuals from Iran being tetraploid while the Turkish and Cypriot ones are diploid. A network analysis showed that the Turkish and Cypriot alleles and one Iranian allele cluster together. The other Iranian allele however is as distant from this cluster as the sister species *Ae. carneum*, suggesting tetraploid origin. Genome wide diversity analyses showed that from the genes that are under selection, 14% have organeller functions largely within mitochondria and/or chloroplasts. As selection happens at the level of traits, we analysed glucosinolate (GS) defence compounds between the populations, showing a low diversity of indolic GS in the Iranian lines while the Turkish/Cypriot lines had a higher indolic concentration and diversity. We found that the *Ae. arabicum* homologs of the tandem duplicated *GSTF9/10* are under selection in the Turkish/Cypriot lines. *GSTF9* and *GSTF10* are involved in the biosynthesis of GS indolic compounds and respond to pathogens and drought stress.

Flow-cytometry and genetic analyses of *Ae. arabicum* showed that the Iranian individuals studied here were tetraploid and had a low level of indolic glucosinolate compounds, while the Turkish and Cypriot lines were diploid and had a high and more diverse level of indolic glucosinolates. In addition to the ploidy differences between the Iranian and Turkish/Cypriot lines, we also observed morphological differences between these populations (personal observations). The Iranian lines have rounder leaves compared to the Turkish and Cypriot lines. Not all morphological characters differ between Iranian and Turkish/Cypriot individuals. With only four leaves until flowering, the Cypriot line is a much faster flowering accession than the Iranian and Turkish lines which have around 9 leaves at flowering (Nguyen *et al.* in prep, personal observation). Future QTL analyses could help identify the genetic basis of variation in flowering time and different GS phenotypes (Nguyen *et al.* in prep, Mohammadin *et al.*, in prep).

The high heterozygosity levels of the tetraploid Iranian lines are consistent with a hypothesis of allopolyploidy. This hypothesis is confirmed by the gene-network analyses, where one allele is incorporated into the Turkish/Cypriot cluster and the other allele is as distant as the outgroup species *Ae. carneum* (Fig. 1 and Suppl Fig 2 and 3). The assumption made for the gene alignments, phased and unphased, is that one of the alleles is more similar to the *Ae. arabicum* reference genome and the other contains the alternative genotypes, hence we could not detect genetic distance(s) between the parental lines. A variable ploidy level within a plant species has been documented in numerous angiosperm families, e.g. *Andropogon gerardi* (Poaceae) (Keeler 1992) and *Artemisia incana* (Asteraceae) (Dolatyari *et al.* 2013). The Brassicaceae core-group also contains many species with different ploidy levels: *Cardamine yezoensis* from eastern Asia (Marhold *et al.* 2010), many *Draba* species (Brochmann 1993, Jordon-Thaden *et al.* 2013), various *Boechera* species (Schrantz *et al.* 2005, Sharbel *et al.* 2005), the European *Biscutella laevigata* (Tremetsberger *et al.* 2002), *Allysum montanum*, *Allysum repens* and *Arabidopsis arenosa* (Španiel *et al.* 2011, Wright *et al.* 2014) all have populations with different ploidy levels. A distinct geographical pattern of ploidy levels can be due to ecological adaptations to small differences in environmental conditions. An example of a distinct pattern of occurrence between diploid and polyploid individuals is from the subspecies complex of *Arabidopsis neglecta* subsp. *neglecta* (diploid)

and *Arabidopsis neglecta* subsp. *robusta* (tetraploid). Here the diploid is found above the tree-line in high alpine habitats and the tetraploid occurs below tree-line in different mountain ranges (Schmickl *et al.* 2012).

A genome-wide analysis of the *Ae. arabicum* genetic diversity showed a higher level of synonymous diversity within the Turkish lines and low levels of genetic diversity for the Iranian lines (Fig. 2). While the Iranian lines were sampled from a very limited geographic area, the distribution of the Turkish lines was over a much larger geographical area. The synonymous diversity levels of the two *Ae. arabicum* clusters are comparable with that of *A. thaliana*, *A. lyrata*, *Boechera stricta* and *Capsella grandiflora* showing similar synonymous substitution rates for self-pollinating Brassicaceae species between 0.003-0.023 (Gossmann *et al.* 2010, Slotte *et al.* 2010, Williamson *et al.* 2014). Similarly, the dN/dS ratios of *Ae. arabicum* fall within the range of dN/dS=0.13-0.21 found in the 257 exonic regions of the annual, biannual and perennial species *Arabidopsis lyrata*, *Capsella grandiflora* and *Noccaea paniculata* (Slotte *et al.* 2010) showing a common background selection pattern.

Although the genetic diversity of the Iranian lines was lower than that of the Turkish lines, there was one exceptional genomic region with an elevated level of synonymous nucleotide diversity for the Iranian lines. This region consisted of 59 genes, mainly coding for basic cellular functions. These 59 genes are all under purifying selection, although this is not significantly different from the rest of the expressed coding genes along the genome. More than 15% of the genes of this Iranian peak are chloroplast chaperonins, belong to the mitochondrion inner membrane or have other functions involved with the organelles. Chaperonins, also called heat shock proteins, are conserved throughout pro- and eukaryotes and are involved in protein folding, especially when cells are stressed (Levy-Rimler *et al.* 2002). They also buffer the destabilization of protein mutations and can as such increase genetic diversity (Tokuriki & Tawfik 2009). Tokuriki and Tawfik (2009) showed that *Escherichia coli* cells without GroEL/GoES chaperons have more than a two-fold decrease of non-synonymous substitution. The GroEL/GoES system is the bacterial homolog of the chaperonins we find here that are under purifying selection.

Glucosinolate (GS) profiles vary depending on tissue and at the species level (Brown *et al.* 2003, Kliebenstein *et al.* 2001). The GS composition of *Arabidopsis thaliana* is locally adapted to its ecological environment (Kliebenstein *et al.* 2001). We found a major difference between the *Ae. arabicum* Iranian and Turkish/Cypriot lines, where the former has only one form of indolic GS and the latter contains four different indolic compounds in much higher concentrations. Plants lacking indolic glucosinolates are more susceptible to necrotrophic fungi (Zhang *et al.* 2015 and the references therein). One *A. thaliana* homolog from the GS biosynthesis pathway under selection in the Turkish/Cypriot lines: *GSTF9* and *GSTF10* involved in indolic GS biosynthesis. Due to our small sample size and the pooled tissues for our transcript isolation we were not able to assess differential gene expression. Although these genes potentially have a role in the formation of the core structure of indolic GS compounds, the diversity of indolic compounds found in our study is mainly due to side chain modifications (Sønderby *et al.* 2010). *GSTF9* and *GSTF10* are tandem repeats belonging to the plant specific 'phi' class of glutathione transferase (Hayes *et al.* 2005, Bergh *et al.* 2016). Glutathione transferases are in general part of the chemical transformation of toxic compounds, catalysing the conjugation of electrophilic centre with tripeptide glutathione

(GSH) (Coleman *et al.* 1997, Frova 2006, Hayes *et al.* 2005). Although *GSTF9/10* are tandem duplicates they seem to be expressed under different stressful conditions in *A. thaliana*. *GSTF9* expression increases in cold treated seedlings, while *GSTF10* expression increases under drought stress (Ryu *et al.* 2009). *GST10* also belongs to a *GST* group that produces proteins in response to increased oxidative stress that is elevated by a pathogen attack (Marrs 1996). Thus the positive selection found here on the *GSTF9/10* paralogs might indicate the adaptation of the Turkish/Cypriot lines to fungal/pathogen attacks or abiotic stresses such as drought stress.

While the genome of *Ae. arabicum* has been used for comparative genetic and genomic studies, little is known about the extant diversity of this species. This is partially due to the current political instability across the Irano-Turanian region, making it hard to collect samples for larger studies. Here the lack of individual and population samples was slightly counter-parted with the large transcriptome datasets generated. This resulted in the finding that the genome-wide genetic diversity was increased within the tetraploid Iranian lines containing genes encoding for chaperonins. Within the Turkish/Cypriot lines, glutathione transferase (GSTs) genes were under selection. Both chaperonins and GSTs are expressed when plant cells are stressed, encouraging the hypothesis that both populations deal with stress-full environments in their own way. A few issues still remain that might be solved with more population samples: whether the diploid and tetraploid individuals occur at distinct locations, whether there has been introgression or allopolyploidy history, and whether the tetraploids are adapted to specific ecological factor(s). Our results are a step forward and present resources that can be used to understand the genetic variation and evolution of *Ae. arabicum*.

### Acknowledgments

This research was funded by the NWO Vernieuwings Impuls VIDI (Grant number: 864.10.001).

### Supplemental files

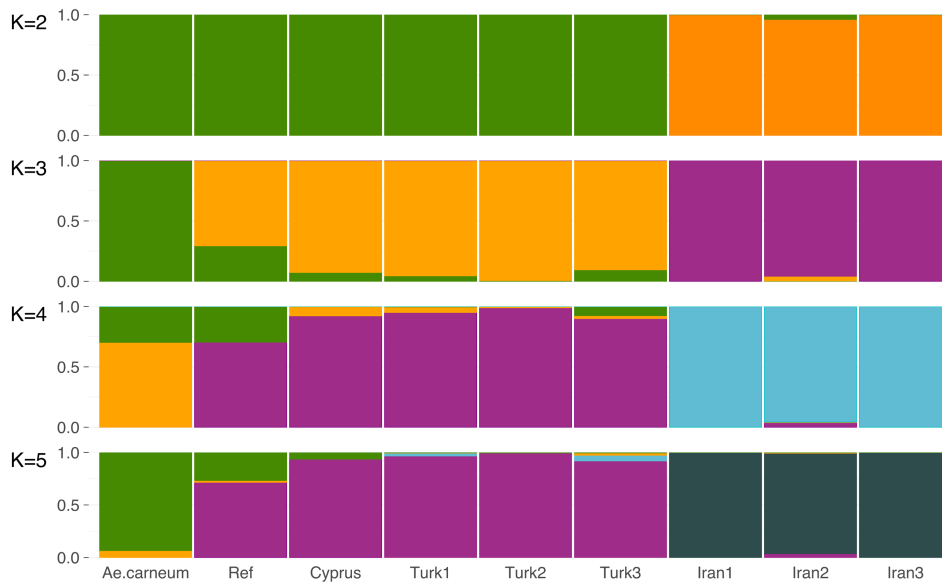
All Supplemental files are available upon request.

### Tables

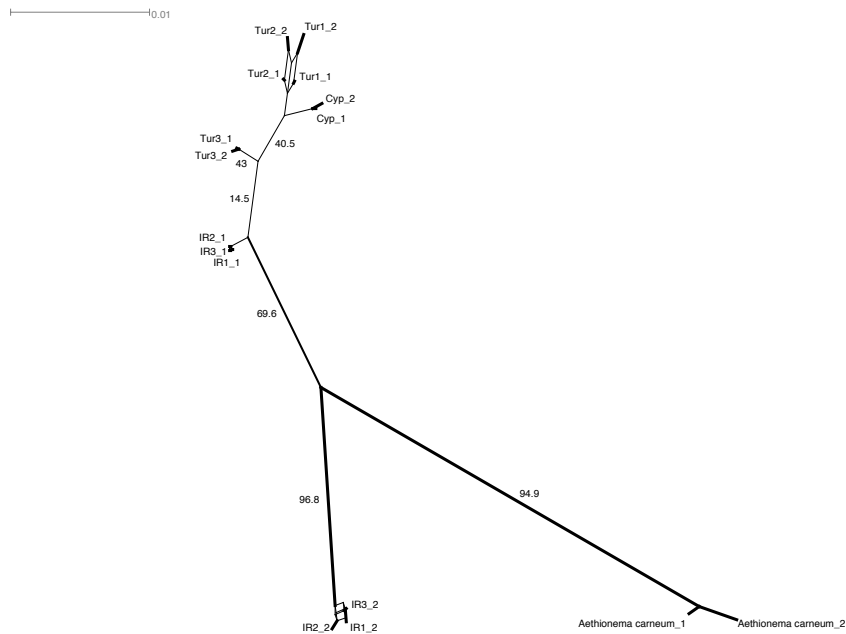
Suppl. Table 1. Population genomic descriptives and function of *Arabidopsis thaliana* homologs of *Aethionema arabicum* genes of the 59 “Iranian peak” genes.

Suppl. Table 2. Population genomic descriptives and function of *Arabidopsis thaliana* homologs of *Aethionema arabicum* glucosinolate genes.

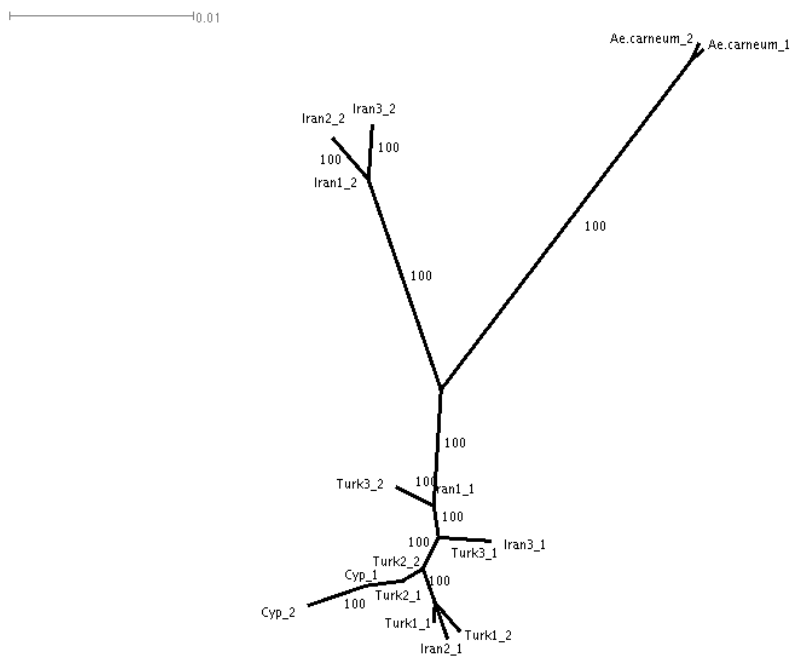
Chapter 4



**Suppl. Fig. 1** STRUCTURE plots showing the clusters of *Aethionema arabicum* populations from Iran (Iran1-3), Turkey (Turk1-3), Cyprus (Cyp), the Reference (Ref) line and *Aethionema carneum* (*Ae. carneum*) transcribed protein-coding genes, with k between 2-5.



**Suppl. Fig. 2** Consensus network of 158 bi-allelic expressed protein coding single copy genes of seven *Aethionema arabicum* lines with *Aethionema carneum* used as the outgroup. The numbers along the branches show the percentage of trees supporting the branch. Naming of the tips is as follow: LocationPopulation\_Allele. Hence Tur1\_1 and Tur1\_2 are the two alleles of the Turk1 population, coming from Turkey. IR= Iran and Cyp= Cyprus.



**Suppl. Fig. 3** Consensus network of 25 bi-allelic expressed protein coding genes from a phased alignment of seven *Aethionema arabicum* lines with *Aethionema carneum* used as the outgroup. The numbers along the branches show the percentage of trees supporting the branch. Naming of the tips is as follow: LocationPopulation\_Allele. Hence Tur1\_1 and Tur1\_2 are the two alleles of the Turk1 population, coming from Turkey. IR= Iran and Cyp= Cyprus.







## Chapter 5

### Major multi-trait quantitative trait locus controls glucosinolate content across developmental stages of *Aethionema arabicum* (Brassicaceae)

Setareh Mohammadin<sup>1</sup>, Thu-Phuong Nguyen<sup>1</sup>,  
M.S. van Weij<sup>1</sup>, Michael Reichelt<sup>2</sup>, M.E. Schranz<sup>1</sup>

Frontiers in Plant Sciences, *under review*

---

<sup>1</sup> Biosystematics, Plant Sciences Group, Wageningen University and Research, Wageningen, The Netherlands

<sup>2</sup> Department of Biochemistry, Max Planck Institute for Chemical Ecology, Jena, Germany

### **Abstract**

The biochemical defence of plants can change during their life-cycle and impact herbivore feeding and plant fitness. The annual species *Aethionema arabicum* belongs to the sister clade to most crucifers of the Brassicaceae core group. Hence it holds a phylogenetically important position to study crucifer trait evolution. Glucosinolates (GS) are essentially Brassicales-specific metabolites involved in plant defence. Using two *Ae. arabicum* accessions (TUR and CYP) we identify substantial differences in glucosinolate profiles and quantities between lines, tissues and developmental stages. We find tissue specific side-chain modifications in aliphatic glucosinolates: methylthioalkyl in leaves, methylsulfinylalkyl in fruits and methylsulfonylalkyl in seeds. We also find large differences in absolute glucosinolate content between the two accessions (up to ten-fold in fruits) that suggest a regulatory factor is involved that is not part of the quintessential glucosinolate biosynthetic pathway. Consistent with this hypothesis, we identified a single major multi-trait QTL (Quantitative Trait Locus) controlling total GS concentration across tissues in a Recombinant Inbred Line (RIL) population derived from TUR and CYP. With fine-mapping, we narrowed the interval to a 58kb region containing fifteen genes, but lacking any known GS biosynthetic genes. The interval contains homologs of both the sulphate transporter *SULTR2;1* and *FLOWERING LOCUS C (FLC)*. Both loci have diverse functions controlling plant physiological and developmental processes and thus are potential candidates regulating glucosinolate variation across the life-cycle of *Aethionema*.

## Introduction

Plant fitness depends on a plants ability to reach the next generation. Thus, plants must be able to defend themselves from herbivores and pathogens throughout their life cycle; first during vegetative growth, then at the time of flowering and finally during the production of fruits and seeds. Pest pressure and effects on survival and fitness can differ greatly across the growth of the plant (Van Zandt 2007). Therefore, defensive compound quality and quantity can shift and be modified during various developmental stages of the plant (Brown *et al.* 2003). Glucosinolates (GS, i.e. mustard oils) and their associated myrosinase enzymes form a two-component chemical plant defence in the Brassicales, defending against herbivores and pathogens. GS are nitrogen and sulphur-rich plant metabolites that hydrolyse, upon contact with the myrosinase enzyme, to form the herbivore-deterrent compounds nitriles and isothiocyanates (Halkier & Gershenzon 2006, Sønderby *et al.* 2010). GS are spatially separated from myrosinase; hence the toxic compounds are only formed after an herbivore attack when the cell is ruptured (Koroleva *et al.*, 2000). The GS biosynthesis additionally has links to fundamental biochemical and developmental processes such as auxin biosynthesis. The Brassicaceae specific IAOX auxin pathway has shared intermediate compounds with indolic GS pathway (Mano & Nemoto 2012). Although all Brassicales contain GS, the highest diversity (of 120 different) GS compounds is found within the economically important family Brassicaceae (Halkier & Gershenzon 2006, Edger *et al.* 2015). This diversity is thought to be due to a combination of gene and genome duplications within Brassicaceae and due to the selective pressure from co-adapting Brassicaceae Pieridae herbivores (Edger *et al.* 2015).

*Arabidopsis thaliana* is used as an important system to understand GS biosynthesis and flowering time. GS quality and quantity change throughout the development of *A. thaliana* (Brown *et al.*, 2003; Petersen *et al.*, 2002) and are influenced by the presence of nutrients, such as sulphur, that are incorporated into GS (Aarabi *et al.* 2016, Falk *et al.* 2007). The availability of these compounds can lead to local adaptation of GS pathway genes (Kliebenstein *et al.*, 2001b). GS are derived from amino acids and can accordingly be divided into three main groups: indolic, aromatic and aliphatic. The molecular mechanisms of the GS biosynthesis pathway have been reviewed and described extensively elsewhere (e.g. Halkier and Gershenzon, 2006 and Sønderby *et al.*, 2010). The diversity of aliphatic GS is, among others, due to different chain lengths and side chain modifications. *BCAT3* and *GS-ELONG* genes regulate chain length during the first step in the GS biosynthesis and *AOP1-3* and *FMO-GS-OX1-5* modify the side chains during the last step of the GS pathway (Sønderby *et al.*, 2010 and the references therein). *Arabidopsis thaliana* is also an important model species for understanding the molecular mechanisms regulating flowering time (Bouché *et al.*, 2016 and the references therein). The switch from vegetative to a generative state is one of the most important moments in the life-cycle of a plant, moderated by abiotic and biotic cues. A plant needs to defend its vegetative and its new valuable generative tissues against potentially shifting herbivore attacks. Hence, shifts between plant development and plant defence traits can be critical to plant fitness. Jensen *et al.* (2015) found a link between the GS pathway and the flowering-time. They incorporated the aliphatic side-chain modifiers, *GS-AOP* genes, in a *AOP-0* background and found that they changed the flowering time of *A. thaliana*. Whether shifts in GS profiles and life-history transitions are seen in other Brassicaceae could establish that this is a general feature of crucifer evolution.

The annual *Aethionema arabicum* belongs to the sister of the Brassicaceae core and hence is at an important position for genetic and genomic comparisons of trait evolution (Schranz *et al.* 2012). *Aethionema arabicum* occurs on steep stony slopes mainly in Iran and Turkey, although populations have also been found more distantly such as in Cyprus and Bulgaria (Velchev 2015). Populations of *Ae. arabicum* go through their entire life cycle between April and June, just before the summer heat strikes (Bibalani 2012), although their flowering time varies throughout the species distribution (Nguyen *et al.* in prep, Mohammadin *et al.* in prep). *Aethionema arabicum* is being studied to understand the molecular mechanisms of fruit heteromorphism (Lenser *et al.* 2016). Heteromorphism is when dehiscent and indehiscent fruits occur on the same infructescence and is hypothesized to be a survival strategy in unpredictable environments. The completion of the *Ae. arabicum* genome has made it possible to study and show the synteny of GS genes between *A. thaliana* and *Ae. arabicum* (Haudry *et al.* 2013, Hofberger *et al.* 2013). However, the diversity of GS profiles in *Ae. arabicum* is not yet known.

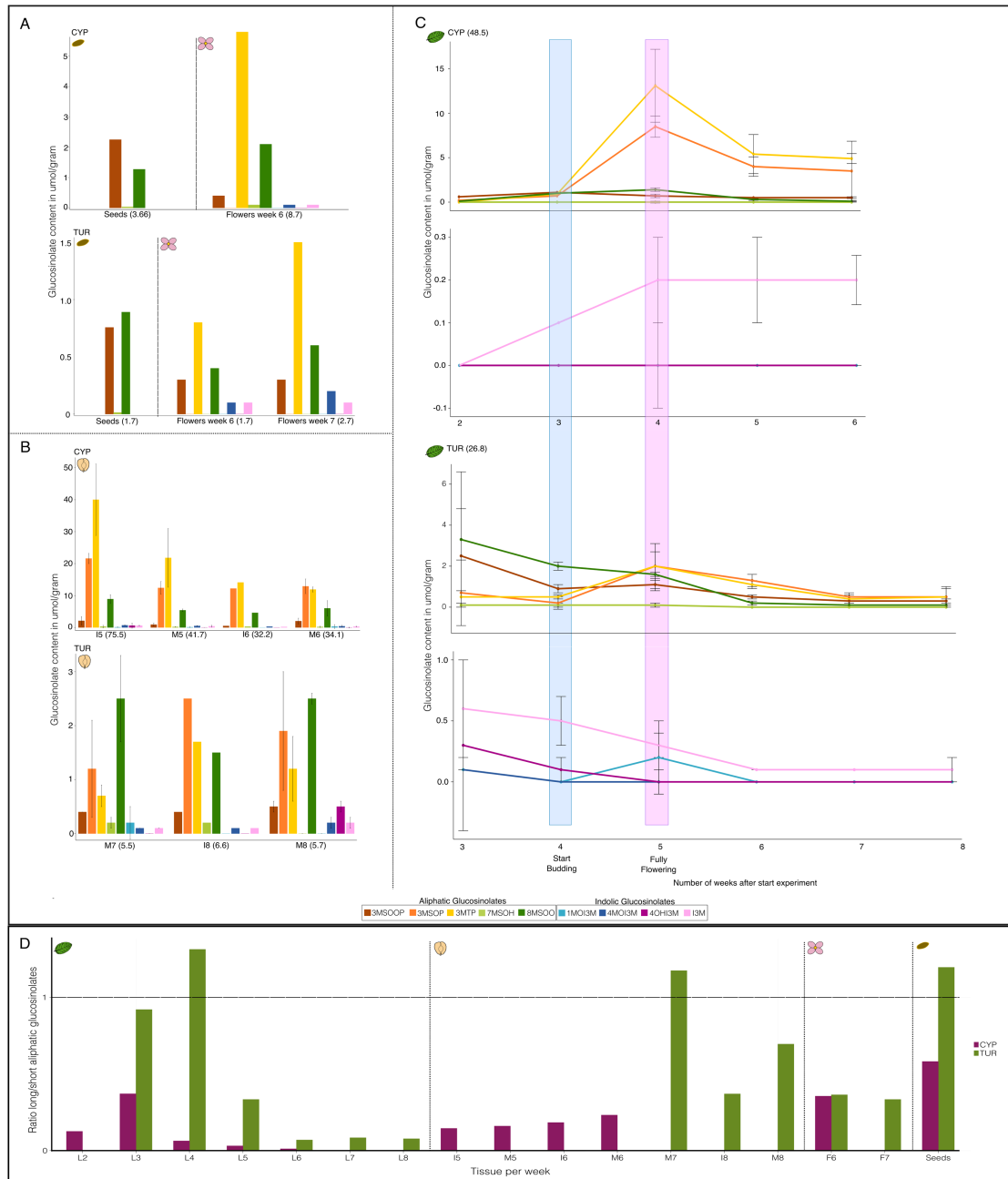
Here we describe the GS content of different tissues and at different developmental stages during a time course. We used two different *Ae. arabicum* accessions (TUR and CYP) differing in their life histories (Nguyen *et al.* in prep). We found that GS content in the leaves depends on the plants developmental stage. Moreover, we found that GS side-chain modifications are tissue-specific and that the two *Ae. arabicum* accessions had a ten-fold difference of GS concentration in the fruits and a two-fold difference in the leaves. We identified the genomic locations controlling GS profiles in *Ae. arabicum* by multi-trait and multi-environment QTL analyses using RIL populations developed from TUR and CYP. In total, we found five QTLs including one major QTL. Two QTL intervals contain homologs of *BCAT3* and *MYB28*, involved in the regulation of long-chained GS formation (Sønderby *et al.* 2010). Although none of the GS pathway homologs were located under our major QTL, this interval contains homologs of the sulphate transporter *SULTR1:2* and the flowering time regulator *FLC*. We believe that *Ae. arabicum* is a good system to understand the link between development and defence and that our study is a valuable first step in that direction.

## Materials and Methods

### Plant material

To measure GS profiles and quantities of *Ae. arabicum* during development, we measured the GS content in seeds, leaves, flowers and fruits of the *Ae. arabicum* lines CYP and TUR originating from Cyprus and Turkey. The description of these lines can be found in Nguyen *et al.* (in prep.). These lines vary in flowering time and occur in very different habitats within the species range of *Ae. arabicum* (Nguyen *et al.* in prep.). CYP and TUR are the parents of the F8 recombinant inbred lines (RILs). We measured GS content in seeds (110 RILs), infructescences and leaves (99 RILs) for QTL analyses.

For germination, seeds were placed on a wet filter paper (demi-water) in a Petri dish and sealed with Parafilm. To measure GS content through development for CYP and TUR seeds were imbibed at 18°C. Seeds showing a radical after imbibition were sown directly in 12cm pots, with five in each pot. Seeds for the QTL experiments were stratified at ~ 4°C after which they were incubated at 18°C to allow germination. Seedlings were individually sown



**Fig. 1** Time course of glucosinolate (GS) composition of different tissues of *Aethionema arabicum* accessions CYP and TUR. Shown are the averages and standard deviations. Bars and points with standard deviations have a sample size of  $n=3$ . Bars and points without standard deviations were from pooled tissue (hence  $n=1$ ). All y-axes have a different scale. Numbers in brackets are total GS content in  $\mu\text{mol}/\text{gram}$  dry weight. A) Seed and flower GS. I = immature fruit, M=Mature fruit. Numbers of the abbreviations along the x-axis are the weeks from the start of the experiment. B) Fruit GS. C) Leaf GS: for every line (CYP or TUR) the top graph are the aliphatic GS and the bottom the indolic GS. Highlighted are the number of weeks after which plants start budding (blue) or are in full bloom (pink). D) Ratio of long versus short aliphatic glucosinolates. The ratio is calculated as the sum of all C7 and C8 GS divided by all C3 GS. The x-axis shows the tissues (L=leaves, I=immature fruits, M= mature fruits, F=flower, Seeds=ripened seeds). Numbers following the abbreviations are the weeks after the start of the experiment. Vertical dotted lines divide the graph into the different tissues.

in 10cm pots. Both experiments were conducted in the climate controlled greenhouse, at Wageningen UR with long day conditions (16h light: 8h dark).

To measure GSs during development, parental plants were sampled weekly from the start of the experiment up to eight weeks. *Aethionema arabicum* does not have a rosette; hence cauline leaves of various ages were sampled and pooled to obtain the GS content throughout the plant. Seedlings, leaves, flowers, immature fruits and mature fruits were sampled separately and in triplicate. Sampling was always done at 11:00 AM, taking diurnal GS variation into account (Petersen *et al.* 2002). For the QTL experiment, leaves were collected when the RILs and parental lines had six fully developed leaves. All plants then showed reproductive buds or were fully flowering. Per line, leaves from two individuals were pooled. Reproductive tissues (infructescence, flowers, and fruits) were collected from the main stem of every RIL and parents one month after the start of the experiment. The reproductive tissues are later referred to as fruits. All samples were immediately stored in liquid nitrogen and stored at -80°C. To assess the GS QTL(s) in seeds, we used dry ripened seeds harvested in 2014 from 110 RILs.

Frozen samples were freeze-dried at -40°C for 24 hours. For GS extraction, samples were ground with 3mm glass beads using 2-10mg of material. For the GS measurements through development if there was <5mg of material, samples were pooled from the same tissue of one parent before GS extraction. To analyse the QTL for seed GS ~10mg of seeds from all RILs were used for GS extraction.

### GS extraction and measurements

2-10mg of freeze-dried leaves or 10mg of seeds were extracted with 1 mL of 80% methanol solution (v:v) containing 0.05 mM intact 4-hydroxybenzylglucosinolate as internal standard. Samples were analyzed after desulfation by HPLC-UV as described in (Burow *et al.* 2006) with the following modification of the chromatographic gradient: water (A)-acetonitrile (B) gradient (0-8 min, 10-50% B; 8-8.1 min, 50-100% B; 8.1-10min 100% B and 10.1-13.5 min 10% B; flow 1.0 mL min<sup>-1</sup>) on a Nucleodur Sphinx RP column (250 x 4.6mm, 5µm, Macherey-Nagel, Düren, Germany). GS were quantified by the peak area of the HPLC-UV chromatogram at 229nm. GS content was normalized to the initial dry weight of the sample (µmol/gram).

### Statistical analyses

Statistical analyses of the development through time was not possible, because of limited sample sizes for all tissues at every time point (n=3 or less due to pooling). Despite the lack of sample size, we found interesting patterns of GS change through time.

The untransformed GS contents were used to for QTL analysis. The linkage map contains eleven linkage groups and is based on 746 SNP from genotype by sequencing analysis of 167 RILs (Nguyen *et al.*, in prep).

Genome-wide QTL analyses were done in Genstat (Payne *et al.* 2009) with a step-size of 5 cM. Single-trait, multi-trait and multi-environment QTL analyses were done for every tissue

separately as done by (Wei *et al.* 2014). This pipeline includes a single interval mapping (SIM) followed by a composite interval mapping (CIM) and a final model selection step for the single trait as well as multi-trait analyses. While single-trait QTL analyses infer QTLs per trait, multi-trait QTL analyses take all the traits simultaneously into account making it possible to assess whether a QTL has a significant effect on a trait. With a backward selection of the found QTLs multi-trait analyses can infer the effect and location of the QTL on every trait (Wei *et al.*, 2014 and the references therein). A multi-environment linkage analysis works in a similar way as a multi-trait analysis, but now the effect of the environment (here the different tissues) on the QTLs per compound is assessed. R/qtl (Broman *et al.* 2003) was used to assess an interval of 1.5 LOD expanded to the markers to assess the genes underlying the QTL. R/qtl makes it possible to include bootstrapping to get the 1.5LOD confidence interval. The significant QTLs were named according to their linkage group, followed by a number depending on the QTL location (e.g. Q1.2 would be the second QTL found on linkage group 1).

Hofberger *et al.* (2013) assessed the homology of the GS pathway between *A. thaliana* and the *Ae. arabicum* v1 genome. We used SynFind (Lyons & Freeling 2008) and the *Ae. arabicum* v2.5 genome in CoGe (Nguyen *et al.* in prep.; Lyons & Freeling, 2008) to establish if any of the homologs found by (Hofberger *et al.* 2013) of the GS pathway occur between a one marker-interval from our QTLs. We used WUblast from the Arabidopsis information resource (Huala *et al.* 2001) (with a significance cut-off value  $\geq e^{-10}$ , including introns and UTRs) to confirm the location of homologs between *Ae. arabicum* and *A. thaliana* for our major QTL. Moreover, we used the transcriptomes of CYP and TUR from Mohammadin *et al.* (in prep) to assess whether the genes under major QTLs were expressed and contained Single Nucleotide Polymorphisms (SNPs). These are transcriptomes from pooled tissues and developmental stages, varying from seed to leaves, and from seedlings to adult plants. SNP quality cut-off value was set as GQ  $\geq 40$  from the variant calls.

## Results

### Glucosinolates during plant development in *Aethionema arabicum*

To investigate if GS contents change through *Ae. arabicum* plant development (temporal) and to assess the GS composition in different tissues (spatial) we measured the GS compounds of *Ae. arabicum* for the parental lines CYP and TUR from a seed to a fully generative (including fruits) stage.

There were nine different GS compounds detected in *Ae. arabicum* (Fig. 1). *Aethionema arabicum* seeds contain only three GS that are all aliphatic and derived from the amino acid methionine: 3MSOOP, 7MSOH and 8MSOO (3-methylsulfonylpropyl, 7-methylsulfinylheptyl and 8-methylsulfinyloctyl glucosinolates respectively, Fig. 1). 8MSOO was the only compound that also occurred in flowers, fruits, and leaves. In addition to 8MSOO *Ae. arabicum* leaves and fruits also contained the Met-derived 3MSOP, 3MTP (3-methylsulfinylpropyl and 3-methylthiopropyl glucosinolates) and the indolic tryptophane-derived 1MOI3M, 4MOI3M, 4OHI3M and I3M (indolyl-3-methyl, 4-hydroxy-indolyl-3-methyl, 4-methoxy-indolyl-3-methyl, 1-methoxy-indolyl-3-methyl glucosinolates respectively, Fig. 1). Thus, compounds differ both in chain-length elongation and in side-chain modifications with a different number of oxygen and sulphur atoms creating sulfinylalkyls, sulfonylalkyls and

**Table 1** Significant QTLs from single trait analyses in *Aethionema arabicum* TURxCYP recombinant inbred lines.

Tissue	GS <sup>a</sup>	QTL <sup>b</sup>	Marker	Pos. (cM) <sup>c</sup>	Low-Upper <sup>d</sup>	-Log <sub>10</sub> (p)	AE <sup>e</sup>	SE <sup>f</sup>	PVE (%) <sup>g</sup>
Leaf	3MSOP	Q8.2	S44_973817	151.0	131.92-167.72	4.57	0.55	0.12	16.41
Leaf	3MTP	Q8.2	S44_827783	153.9	143.57-164.25	7.99	1.99	0.317	28.34*
Leaf	8MSOO	Q6.2	S40_550522	78.8	61.78-95.83	6.26	0.40	0.075	19.44
Leaf	8MSOO	Q8.2	S44_609479	158.1	121.15-167.72	4.36	0.32	0.075	12.44
Leaf	A/I	Q10.1	S61_1993061	124.5	85.19-160.54	3.54	5.73	1.52	12.12
Leaf	A/I	Q8.2	S44_827783	153.9	123.08-167.72	4.28	6.06	1.43	13.55
Leaf	All GS	Q8.2	S44_609479	158.1	147.14-167.72	7.5	2.95	0.49	27.13*
Fruit	3MSOP	Q8.2	S44_973817	151.0	140.99-160.92	8.26	2.98	0.46	30.74*
Fruit	3MTP	Q6.2	S58_8426	72.8	30.92-114.76	4.48	1.78	0.41	12.43
Fruit	3MTP	Q8.2	S44_827783	153.9	139.37-167.72	7.3	2.42	0.41	22.92
Fruit	8MSOO	Q8.2	S44_827783	153.9	145.97-161.85	10.36	3.34	0.448	37.12*
Fruit	4OH13M	Q5.1	S53_4972590	25.0	0-71.94	4.07	0.42	0.10	11.94
Fruit	4OH13M	Q8.2	S44_827783	153.9	141.47-166.35	7.62	0.61	0.10	25.78*
Fruit	All GS	Q6.2	S58_8426	72.8	0-129.44	4.64	4.4	0.99	10.28
Fruit	All GS	Q8.2	S44_827783	153.9	145.39-162.43	11.68	8.12	0.99	34.99*
Seed	3MSOOP	Q8.1	S81_435439	7.9	0.73-58.6	7.89	3.74	0.871	10.53
Seed	3MSOOP	Q8.2	S44_827783	153.9	143.48-164.34	9.34	5.96	0.87	26.81*
Seed	7MSOH	Q6.1	S5_745413	16.6	0.0-41.18	5.62	0.15	0.03	14.6
Seed	7MSOH	Q8.2	S44_973817	151	139.57-162.35	8.81	0.2	0.03	24.99*
Seed	8MSOO	Q8.2	S44_973817	151	140.04-161.88	7.57	6.37	1.06	25.84*
Seed	All GS	Q8.2	S44_973817	153.9	146-161.83	10.17	33.7	1.79	13.03

<sup>a</sup>) Glucosinolate Compound. A/I= ratio Aliphatic by Indolic. <sup>b</sup>) QTL name. First number corresponds to Linkage Group. <sup>c</sup>) Position along linkage group in centimorgan (cM). <sup>d</sup>) Lower and upper bound of QTL from Genstat. <sup>e</sup>) Additive effect. Negative effects are from CYP, positive from TUR. Negative or positive effect means that the CYP or TUR allele has a stronger effect. <sup>f</sup>) Standard error of additive effect. <sup>g</sup>) Percentage of explained variance. \*) major QTL (PVE ≥ 25%, after Burke et al 2002).



thioalkyls for aliphatic GS and adding methoxy-groups to the indolic GS. The greatest variety of compounds was found in the early developing leaves and fruits of TUR (Fig. 1B and C), although this variation decreased through time in the leaves.

The GS quality and quantity changed through time and varied between tissues (Fig. 1, Suppl. Table 1). In all tissues the ratio of long- vs. short-chained aliphatic GS was skewed towards the short-chain compounds (Fig.1D). GS content decreases after fruiting sets in in both CYP and TUR (Fig. 1C and Suppl. Fig. 1). Moreover, CYP shows an increase of aliphatic GS after budding that peaks during flowering (Fig 1C, Suppl. Fig. 1). It is not clear whether GS also increase in TUR after budding, as we do not have any data for the early seedlings of TUR (week 2). 3MTP and 3MSOP increase in CYP plants towards 13.1  $\mu\text{mol}/\text{gram}$  ( $\pm 4.1$   $\mu\text{mol}/\text{gram}$  SD) and 8.5  $\mu\text{mol}/\text{gram}$  ( $\pm 1.2$   $\mu\text{mol}/\text{gram}$  SD), respectively. The indolic I3M rises to a level of 0.2  $\mu\text{mol}/\text{gram}$  ( $\pm 0.1$   $\mu\text{mol}/\text{gram}$  SD) when CYP is fully flowering and levels off at 0.2  $\mu\text{mol}/\text{gram}$  for the rest of the lifecycle. In the TUR leaves 3MTP and 3MSOP both increase to a level of 2.0  $\mu\text{mol}/\text{gram}$  each (3MTP  $\pm 1.1$   $\mu\text{mol}/\text{gram}$  SD and 3MSOP  $\pm 0.7$   $\mu\text{mol}/\text{gram}$  SD) during flowering. While the short chain GS increase, the long chain 8MSOO decreases in the TUR lines from 3.3  $\mu\text{mol}/\text{gram}$  ( $\pm 3.3$   $\mu\text{mol}/\text{gram}$  SD) a week before bolting to 0.1  $\mu\text{mol}/\text{gram}$  ( $\pm 0$   $\mu\text{mol}/\text{gram}$  SD) three weeks after flowering (Fig. 1).

In almost all sampled tissue CYP has a higher GS content than TUR. The only exception to this are the leaf indolic GS, where TUR starts on average with a higher indolic GS content than CYP (Fig. 1C; CYP = 0  $\mu\text{mol}/\text{gram}$  for all indolic GS and TUR 1MOI3M=0.1  $\mu\text{mol}/\text{gram}$ ; 4MOI3M=0.1  $\mu\text{mol}/\text{gram}$ ; 4OHI3M=0.3  $\mu\text{mol}/\text{gram}$ ; I3M=0.6  $\mu\text{mol}/\text{gram}$ ). Although the TUR indolic GS decrease towards 0.1  $\mu\text{mol}/\text{gram}$  or even less, the CYP indolic I3M increases through time.

### QTL analyses

To understand the genetic regulation of GS in *Ae. arabicum* we investigated the GS profiles and identified the genomic locations underlying the GS phenotype in *Ae. arabicum* leaves, fruits and seeds from RILs and their parental lines TUR and CYP. In addition to single-trait QTL analyses we also applied multi-trait and multi-environment QTL analyses to assess the effects of the QTLs on the different compounds and on the different tissues.

The leaf and fruit samples of the RILs were taken after budding or even during flowering and contain only 3MSOP, 3MTP, 8MSOO, 4OHI3M and I3M. The segregation spectrum of the RILs was similar for all GS whether they were isolated from leaves, fruits or seeds (Suppl. Fig. 2). However, the GS concentrations depended on compound as well as tissue, with indolics being lower than aliphatics (Suppl. Fig. 2).

For the single-trait single-environment analysis, we found six different QTLs on 4 different linkage groups Q5.1 on LG5, Q6.1 and Q6.2 on LG6, Q8.1 and Q8.2 on LG8 and Q10.1 on LG10 (Table 1). Three of the QTLs (Q6.1, Q6.2 and Q8.2) also occur in the multi-trait and multi-environment QTL analyses (Table 2 and Suppl. Table 2). Q8.2 is a major QTL throughout all our analyses. The single-trait single-environment analysis shows a QTL for the indolic 4OHI3M. However, this peak is not present in the multi-environment QTL analysis (Suppl. Table 2).

The multi-trait single environment analysis shows that 4OHI3M is only significantly affected by the QTLs within the fruits (Fig. 2). The indolic I3M significantly interacts with Q8.2 in the leaves (Fig. 2). The multi-trait QTL (Fig. 2) shows that Q8.2 always interacts with all the aliphatic GS in all tissues. There is a difference in interaction between 3MSOP in leaves and fruits: while 3MSOP has a significant interaction with both QTLs in fruits there is only the significant interaction with Q8.2 in leaves (Fig. 2). 3MTP however has a significant interaction with all the QTLs in leaves. Hence the occurrence of 3MTP or 3MSOP seems to be tissue specific. Moreover, compared to leaves and fruits, seeds have two unique loci: Q2.1 and Q8.1.

The multi-trait (Fig. 2) and multi-environment (Suppl. Table 2) analyses both show a large effect from the TUR allele for Q6.2 and Q8.2 in leaves and seeds, and for Q6.2 in leaves and fruits. However, Q2.1 in seeds (for 7MSOH and 8MSOO) had a high effect from the CYP allele. Q8.2 had a strong effect from the CYP allele for the indolic GS, which was also the case for leaf Q6.1 (Fig. 2). The combination of the low GS levels for TUR compared to CYP and the strong effect of the TUR allele on the QTLs suggests that our QTLs are inhibitors of GS synthesis or transport.

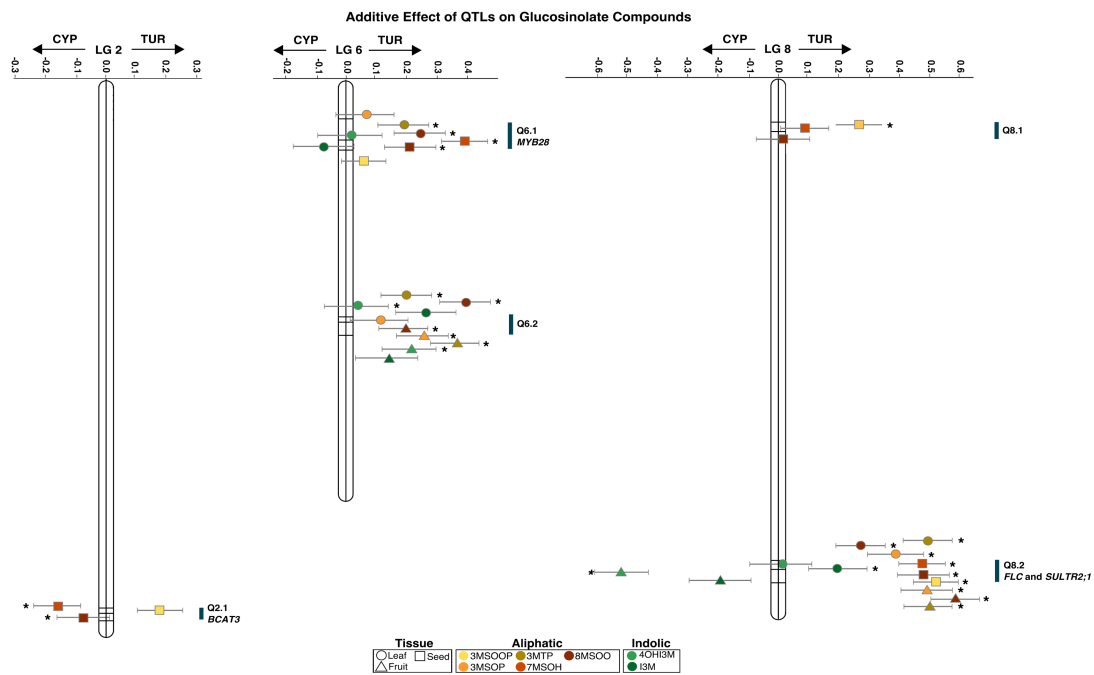
As leaves and fruits contain the same compounds, we used a multi-environment QTL analyses to assess the effect of the QTLs on the tissues for every compound. In addition to the already shown QTLs of the single trait single environment analysis and the multi-trait single environment analysis, the multi-environment single trait analysis has three new QTLs: Q1.1, Q3.1 and Q8.3 (Suppl. Table 2). This comparison shows that the QTLs are significantly correlated mainly in fruits. An exception to this is 3MTP where both leaves and fruits are significantly correlated with the QTLs.

Using the homology and synteny between *A. thaliana* and *Ae. arabicum* of GS pathway genes (Hofberger *et al.* 2013) we assessed whether any of the genes were potential candidate genes for any of our QTLs. We found *BCAT3* was within the confidence interval of Q2.1 and *MYB28* was in the confidence interval of Q6.1. None of the GS pathway genes were coded by the *Ae. arabicum* genes under the major QTL, Q8.2, that appears in every comparison. Using WU-BLAST (Huala *et al.*, 2001; Suppl. Table 3) we identified a total of 87 genes within the confidence interval. Using the raw genotype information from Nguyen *et al.* (in prep.) we were able to define more precisely the region to 58kb and fifteen genes (Suppl. Table 3). The genes have diverse functions including fatty acids biosynthesis, ethylene-activated signalling, proteolysis, Pollen Ole e1, one unknown protein, the sulphate transporter *SULTR2;1* and the *FLOWERING TIME LOCUS C*.

Eight out of the 15 genes within the 58kb interval had SNPs within the transcriptomes of Mohammadin *et al.* (in prep; Suppl. Table 3). Although *SULTR2;1* has SNPs in this transcriptome dataset the SNP seen in *FLC* did not seem to have a polymorphism according to our cut-off values (there was one SNP in *FLC* with GQ=39, our cut-off was GQ ≥ 40).

## Discussion

Here we present the correlation between the reproductive phase change and the composition of defence compounds in the annual Brassicaceae *Aethionema arabicum*. Although the



**Fig. 2** Additive effect of QTLs from a multi-trait QTL analysis for different glucosinolates (GS) from different tissues of *Aethionema arabicum*. Shown are the Linkage Groups (numbers above the x-axis) with lines in them for the QTL. The middle line is always the main locus; the other two are its closest markers. Positive effects are from the TUR allele, and negative values from the CYP allele. Points represent the tissues (different shapes) and GS compounds (different colours, see legend at bottom of figure) with their standard error (grey whiskers). QTL with a significant ( $p < 0.05$ ) effect on the GS are denoted with an asterisk. Blue bars along the linkage groups represent the QTL position and candidate genes. All points belonging to the same QTL are off-set for visibility.

GS pathway has been extensively studied in *Arabidopsis thaliana*, the information from a phylogenetically distant crucifer may elucidate alternative GS regulation. We show that the major genomic location (Q8.2) associating with GS variation contains fifteen genes, among which are the sulphur transporter *SULTR2;1* and the *FLOWERING TIME LOCUS C (FLC)*, genes that are not known to be directly involved in the GS biosynthesis pathway.

*Aethionema arabicum* ripened seeds, from CYP as well as from TUR, have lower GS diversity than *Ae. arabicum* fresh leaves. This differs from other crucifers, e.g. *A. thaliana*, *Brassica oleracea* and *B. napus*, where the GS diversity and concentration are the highest in the seeds and decrease in the following order in the inflorescence, siliques, leaves and roots (Brown *et al.*, 2003; Sotelo *et al.*, 2014; Velasco *et al.*, 2008). *Aethionema arabicum* fresh fruits, including seeds, have very high GS levels (Fig. 1) comparable to the levels found in *A. thaliana* seeds (Brown *et al.* 2003). While all *Ae. arabicum* tissues contain indolic GS, their ripened seeds lack these compounds. The difference in indolic GS is also seen between the seeds and leaves of *A. thaliana*, *B. oleracea* and *B. napus* (Brown *et al.* 2003, Petersen *et al.* 2002, Sotelo *et al.* 2014, Velasco *et al.* 2008, Kliebenstein *et al.* 2001). Aliphatic GS are known to have a negative effect on the survival and growth of herbivorous insects (Beekweelder *et al.* 2008) explaining the persistence of aliphatic GS in the two-year-old *Ae. arabicum* seeds, but also the presence of GS in seeds of Brassicaceae that lack GS in their vegetative tissue (Windsor *et al.* 2005). The difference between young versus old tissue might indicate a breakdown process of (indolic) GS in the seed cells over time or a

translocation process (away from the seeds) as the seeds ripen, though it might also be correlated to the developmental stage of ripened seeds, hypotheses still to be tested.

The two parental ecotypes used here (CYP and TUR) had a ten-fold difference in GS concentrations in the fruits and two-fold difference concentration in leaves, with CYP always having the higher aliphatic GS concentration (Fig. 1). These differences seem to reflect the extremes found in other Brassicaceae species. For example Italian horseradish (*Armoracia rusticana*) roots can differ up to twenty times in GS concentration depending on the accession (Agneta *et al.* 2014); American wild radish (*Raphanus raphanistrum*) accessions has more than 20x difference in GS concentrations in its secondary branches of plants from North Carolina vs Mississippi (Malik *et al.* 2010); and *A. thaliana* leaves can have extremes of more than 10x differences in total GS concentrations (Kliebenstein *et al.* 2001).

The different tissues of *Ae. arabicum* all show a skewed ratio of long- versus short-chain GS and different side chain modifications. A similar pattern has been shown in *A. thaliana*, where ecotypes with higher amounts of C3 GS had lower C8 to C7 ratios (Kliebenstein *et al.* 2001). The negative long-short correlation could be a biochemical effect whereby short GS precursors are kept in the chain elongation loop as long as they are not used (Olson-

**Table 2** Significant QTLs from multi trait analyses in *Aethionema arabicum* TURxCYP recombinant inbred lines. The percentage of variability explained is shown for every GS compound that had a significant ( $\alpha \leq 0.05$ ) interaction with the QTL.

QTL	Marker	Tissue	LG <sup>a</sup>	Pos. (cM) <sup>b</sup>	-Log(p)	GS <sup>c</sup>	PVE % <sup>d</sup>
Q6.1	S5_745413	Leaf	6	16.57	4.34	3MTP	3.6
						8MSOO	5.9
Q6.2	S40_63849	Leaf	6	74.62	5.66	3MTP	4.0
						8MSOO	15.5
						I3M	6.9
Q8.2	S44_827783	Leaf	8	153.91	9.25	3MSOP	14.9
						3MTP	24.3
						8MSOO	7.4
						I3M	3.8
Q6.2	S40_63849	Fruit	6	74.62	5.05	3MSOP	6.4
						3MTP	12.9
						8MSOO	3.6
						4OHI3M	4.3
Q8.2	S44_827783	Fruit	8	153.91	17.26	3MSOP	24.0
						3MTP	24.4
						8MSOO	34.1*
						4OHI3M	26.8*
Q6.1	S5_745413	Seed	6	16.57	9.28	7MSOH	16.8
						8MSOO	6.6
Q2.1	S13_476613	Seed	2	167.30	7.56	3MSOOP	3.0
						7MSOH	3.5
Q8.1	S93_383588	Seed	8	11.54	4.14	3MSOOP	7.2
Q8.2	S44_827783	Seed	8	153.91	10.54	3MSOOP	26.6*
						7MSOH	25.7*
						8MSOO	18.5

<sup>a</sup>) Linkage Group. <sup>b</sup>) Marker Position in centimorgans. <sup>c</sup>) Significant glucosinolate compounds ( $\alpha \geq 0.05$ ). <sup>d</sup>) Percentage of variance explained. \*) Major QTL (PVE  $\geq 25\%$ , after Burke *et al.* 2002).

Manning *et al.* 2015). *Aethionema arabicum* has tissue specific oxidization levels of GS side chain modifications: leaves were mainly correlated with 3MTP, while 3MSOP is linked to fruits and 3MSOOP occurs only in the flowers and seeds (Fig. 1), this is also reflected in the multi-trait QTLs (Fig. 2). Side chain oxidation (OHP vs 3MSOP) has a negative effect on the weight gain of herbivores (Rohr *et al.* 2009) presenting an array of testable hypotheses for the ecological effect of highly oxidized GS in *Ae. arabicum* seeds. Knowing the genetic architecture controlling GS variation in *Ae. arabicum* can elucidate GS regulation and tissue specific side chain modifications. Moreover the correlation between the long and short aliphatic GS indicates a similar regulatory factor. However, none of the expected GS pathway genes, e.g. *MAM*'s, *CYP*'s, *GS-OX*'s or *AOP*'s were associated with identified QTLs. Only two of the minor QTLs contain genes involved in the chain elongation: *BCAT3* and *MYB28* (Beekweelder *et al.* 2008, Knill *et al.* 2008). *BCAT3* is involved in the chain elongation process and *BCAT3* knockouts increase the level of long chain GS compounds (Knill *et al.* 2008, Søndery *et al.* 2010). Although it is not known how *MYB28* is involved in the long-chain GS biosynthesis it has been shown that the knockout *myb28* blocks the expression of long-chain GS (Beekweelder *et al.* 2008).

The GS quantity and quality changes throughout the development of *Ae. arabicum*. This is most clearly seen in CYP where at the onset of budding there is an increase in aliphatic GS (Fig. 1 and Suppl. Fig. 1). Moreover both in CYP and TUR the level of GS decreases after flowering (Fig. 1 and Suppl. Fig. 1). Combining the change in GS throughout *Ae. arabicum*'s development and the large difference of GS concentration found between CYP and TUR strongly suggests some regulatory factor(s) other than the known genes involved in the GS biosynthesis pathway. The multi-trait and multi-environment QTL analyses indeed show one major QTL (Q8.2) explaining up to 37% of the variation of the lines (Table 1). Preliminary fine mapping this region indicated fifteen genes, among which the sulphate transporter *SULTR2;1* and *FLOWERING TIME LOCUS C (FLC)*, one of the MADS-box transcription factors that regulated flowering in Brassicaceae (Ietswaart *et al.* 2012, Bouché *et al.* 2016) are the most intriguing candidates.

*SULTR2;1* is involved in the root to shoot sulphate transport (Gigolashvili & Kopriva 2014). Sulphur is an essential macronutrient for plant development (Gigolashvili & Kopriva 2014). Sulphur is used in the biosynthesis of several compounds varying from amino acids to proteins, co-enzymes, vitamins and defence metabolites like GS (Gigolashvili & Kopriva 2014, Falk *et al.* 2007). With at least two sulphur atoms GS can include ~30% of the plants sulphur (Aarabi *et al.* 2016). The addition of sulphur can increase GS levels with 25%-50%, depending on the amount of sulphur and the treatment (Falk *et al.* 2007). There are four groups of sulphate transporters: high affinity transporters (*SULTR1*'s); plastid membrane transporters and low affinity transporters such as (*SULTR2*'s); transporters of the symbiosome membrane of the legume:rhizobia symbiosis (*SULTR3*'s); and transporters with an unknown function (*SULTR4*'s) (Gigolashvili & Kopriva 2014). The low affinity sulphate transporters, such as *SULTR2;1* depend more on sulphur availability and hence respond quicker to sulphur deficiency (Falk *et al.* 2007). Under sulphur deficient circumstances GS are broken down and used as a sulphur source (Falk *et al.* 2007) while the biosynthesis of GS is repressed (Aarabi *et al.* 2016). The importance of sulphur for GS and its locality under our major QTL might indicate an indirect relation between sulphur transport and GS formation, especially in *Ae. arabicum* that originates from steep stony slopes in Irano-Turanian region, where its

developmental time is only between the hot and dry summers and the cold winters. The MADS-Box gene *FLOWERING TIME LOCUS C (FLC)* is expressed throughout the life cycle and in many tissues of Brassicaceae species. However, it is primarily known as a floral repressor in meristems where expression is stably repressed upon prolonged cold-treatment or vernalization (Deng *et al.* 2011, Bouché *et al.* 2016, Ietswaart *et al.* 2012). *Aethionema arabicum* is a relatively fast flowering annual and does not require vernalization. Potentially, *Ae. arabicum FLC* is involved in the regulation of GS biosynthesis. Our data show that the levels of leaf GS increase at the onset of budding in *Ae. arabicum*. The transition from a vegetative to a generative life stage effects many aspects of the plant biology. For example, the biosynthesis of GS can reduce fitness in *A. thaliana* (Kliebenstein 2004). In *A. thaliana* and *B. napus* even the production of a new leaf leads to an increase of GS concentration in the new leaf compared to the older leaves (Brown *et al.* 2003). *FLC* has more than 500 binding sites in the *A. thaliana* genome with *CYP79B3* being one of them (Deng *et al.* 2011). *CYP79B3* belongs to the cytochrome P450 *CYP79* family and is involved in the formation of the core structure of indolic GS (Sønderby *et al.* 2010). The major QTL (Q8.2) found here in *Ae. arabicum* likely indicates a link between GS biosynthesis and development. Jensen *et al.* (2015) showed that the introduction of the *GS-AOP* genes in *AOP-0* lines does not change the GS levels, but influences flowering time. This effect depended on the genetic background and could vary between an increase and decrease of flowering time (Jensen *et al.* 2015). They hypothesized that *AOP2* and *AOP3* could mediate the cross talk between flowering and defence.

Our *Ae. arabicum* lines show a natural transition from vegetative tissue to generative tissue correlates with a transition in GS content, although it could be regulated vice versa. QTL analyses for GS content of three different tissues all point to one major QTL containing two potential candidates *SULTR2;1* and *FLC*. Future gene expression analysis, transformation experiments, fine mapping and phenotyping of more accessions could help to understand whether and if so in which way these genes and traits are involved in the plants defence pathway.

### **Funding**

This work was supported by the grants from NWO Vernieuwings Impuls VIDI (Grant number: 864.10.001).

### **Acknowledgements**

We thank A.M. Houtkooper with his help in fine-tuning the figures.

### **Supplemental files**

All Supplemental files are available upon request.

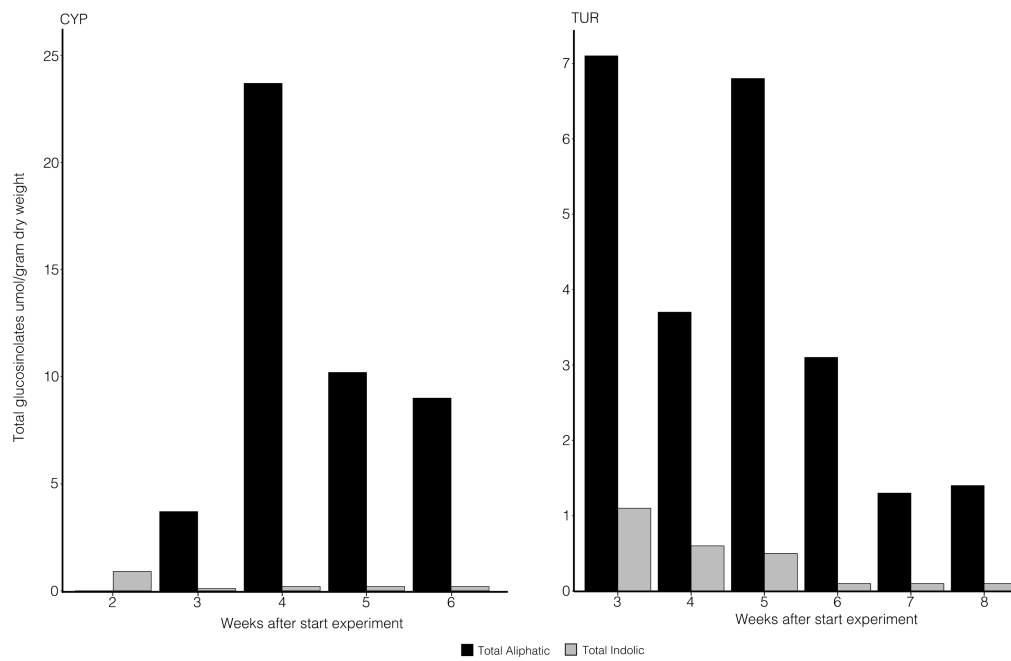
### **Tables**

Suppl. Table 1. Average glucosinolate values of TUR and CYP leaves.

Suppl. Table 2. Significant QTLs from multi environment analyses in *Aethionema arabicum* TURxCYP recombinant inbred lines

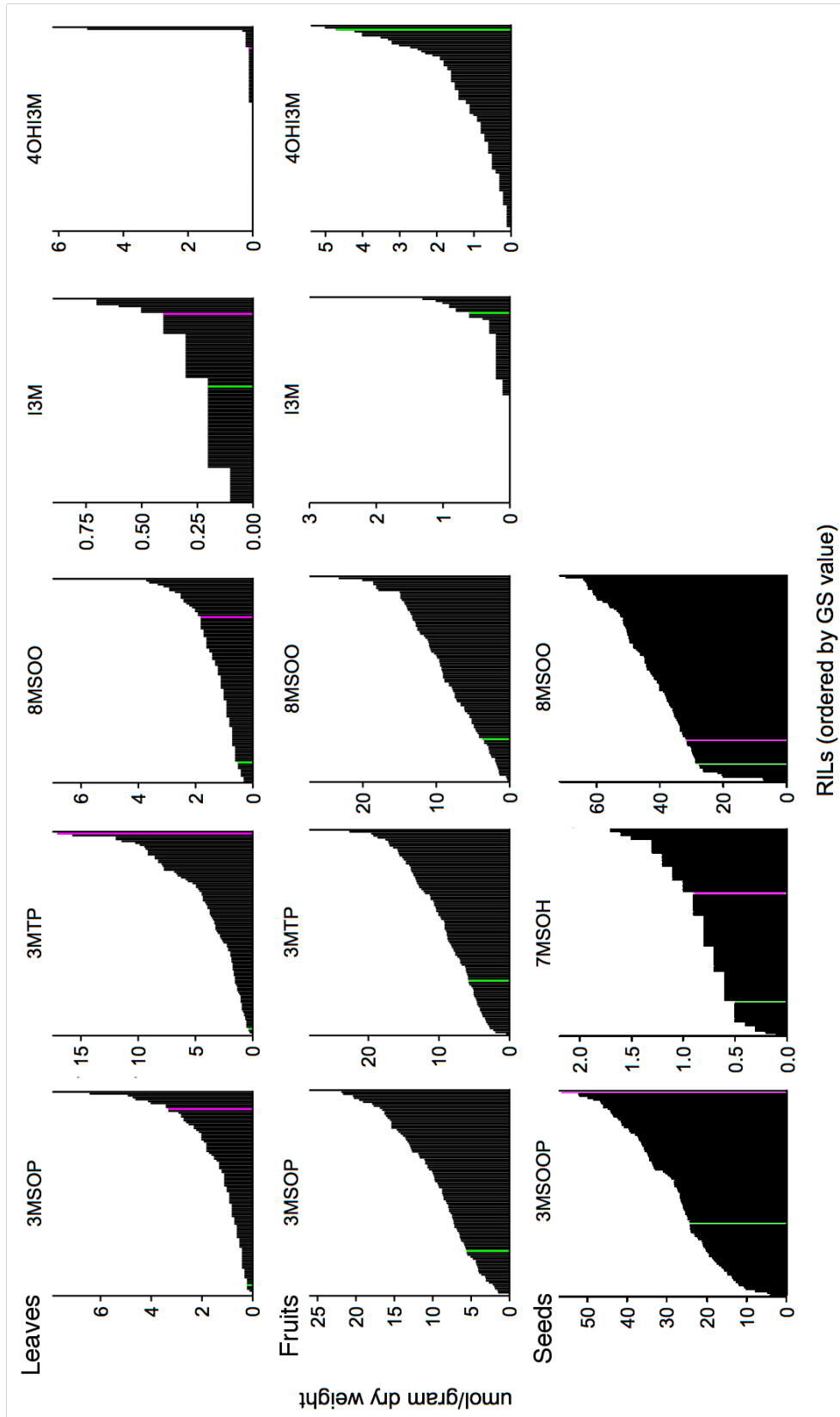
Suppl. Table 3. *Arabidopsis thaliana* homologs of *Aethionema arabicum* genes in the interval of the major Q8.2 and the fine-mapping details.

Chapter 5



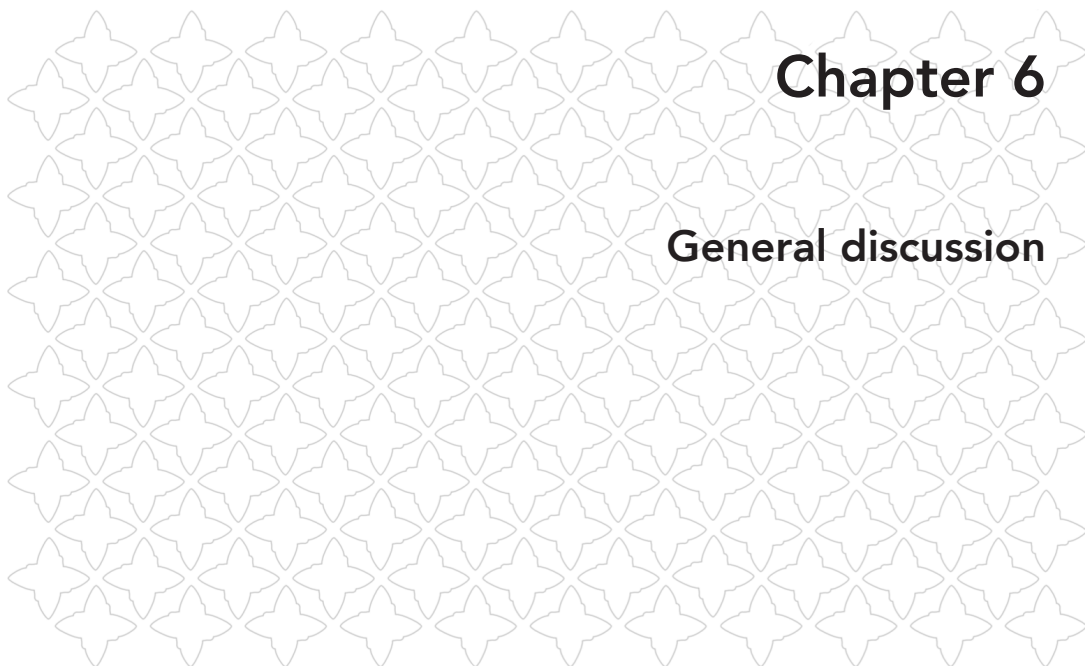
**Suppl. Fig. 1** Total Aliphatic and indolic glucosinolates ( $\mu\text{mol}/\text{gram}$  dry weight) measured in leaves of *Aethionema arabicum* (A for CYP and B for TUR). The x-axis denotes the weeks after the start of the experiment.





Suppl. Fig. 2 Glucosinolates (umol/gram dry weight) per RIL line and the parental lines TUR (green) and CYP (pink) for leaves, fruits and seeds of *Aethionema arabicum*.





# **Chapter 6**

## **General discussion**

The intention of this thesis is to use multiple biological disciplines and approaches to understand the evolution of *Aethionema* (Brassicaceae). In this final chapter is a synthesis of these disciplines. Answers to the questions posed in Chapter 1 are discussed and challenges for future research are highlighted.

### **The effects of whole genome duplications**

Whole Genome Duplications (WGD or polyploidisation) greatly alter the genomic content of a species and thus impact speciation (Soltis *et al.* 2009). Ohno postulated that gene duplications likely have a major role in evolution (Ohno 1970). He stated that it is easier to duplicate something that already exists than to create something new (Ohno 1970, Wolfe 2001). After polyploidisation, duplicated genes can be retained, get a new function (neofunctionalise), subfunctionalise or lose their function by becoming a pseudo-gene (Comai 2006). Moreover, after a WGD event a genome can fractionate and become diploid again: the process of diploidisation (Comai 2006, Barker *et al.* 2009). Diploidisation does not affect all genomic regions in the same degree. This is due to the different rates of molecular evolution between protein coding genes (Wolfe 2001). For example, compared to genes that remain duplicated those that return to single-copy are under severe selection pressure and are biased towards housekeeping functions (De Smet *et al.* 2013).

A WGD event has a severe effect on a genome and it has been regarded as an evolutionary dead-end and/or as a mechanism driving species radiations (Soltis *et al.* 2009, Mayrose *et al.* 2011, Arrigo & Barker 2012). Polyploid individuals seem to have a higher extinction rate than their diploid congeners (Mayrose *et al.* 2011). They have meiotic incompatibility challenges, start with very low population sizes or selection can become inefficient due to the presence of multiple copies of an allele (Arrigo & Barker 2012 and the references therein). There are estimates that 15% of speciation events are due to WGDs (Arrigo & Barker 2012). Traces of ancient polyploidy events, e.g. the gamma WGD event at the basis of the angiosperms (Jiao *et al.* 2011), are found within nearly all plant genomes. Twenty-five percent of the *A. thaliana* genes are retained from ancient polyploidy events (Blanc & Wolfe 2004). This can go up to 67% for *Glycine max* (Schmutz *et al.* 2010). These conserved duplicated gene pairs and the selection pressure on housekeeping genes to be singletons encourages the hypothesis that WGD events are not dead ends, but can lead to the development of novel traits over time (Schranz *et al.* 2012).

The advantages of polyploid individuals seem to be related to severe unstable environments (Kagale *et al.* 2014), when an increase in gene content and expression, increased heterozygosity, neofunctionalisation and epigenetic reprogramming might all be advantageous (Kagale *et al.* 2014, Fawcett *et al.* 2009, Comai 2006). For example, many ancient WGD events coincide at the K-T boundary (Vanneste *et al.* 2014, Kagale *et al.* 2014, Fawcett *et al.* 2009), the major extinction event that caused the extinction of the dinosaurs and the extinction of many plants, i.e. up to ~60% of the North American plant species (Fawcett *et al.* 2009 and the references therein). This probably explains the non-uniform distribution of WGDs: most WGD events are near the tips of the angiosperm phylogeny, when many of the currently known plant families originated, such as the Brassicaceae (Fawcett *et al.* 2009, Mayrose *et al.* 2011, Vanneste *et al.* 2014, Kagale *et al.* 2014).

The diploidisation process following a polyploidisation event causes the re-arrangement of genomic regions. For example, multiple polyploidisation events within the Brassicaceae followed by diploidisation have rearranged the genomic content, although some genomic blocks have been conserved (Lysak *et al.* 2016). Gene function depends on the sequence as well as the position of a gene (Lyons & Freeling 2008). Genomic position of genes and their regulatory elements should hence be taken into account in analyses on functional conservation (Chapter 2). Incorporating positional conservation might help to separate paralogs from orthologs and can elucidate the conservation of presumed functional genomic units (Chapter 2). Incorporating positional conservation of genetic elements can dilute the old definitions on lineage specificity, as genetic elements can be conserved by position, but can lack conservation on the basis of their sequences (Chapter 2). However, incorporating gene collinearity in gene conservation research can elude specific gene networks that share an evolutionary history (Zao *et al.* in prep).

### **From paleo-ecological circumstances to species genetic diversity**

Speciation depends on the genetic and genomic background within a species and on the biotic and abiotic circumstances. Both genomic background and local (a)biotic factors define the level of plasticity of a species. The historical biogeography of *Aethionema* showed us that the major geological events during the Miocene in the Irano-Turanian region correlate with the diversification of three well-supported *Aethionema* clades (Chapter 3). Local geological events caused local climatic differences that probably enforced the speciation of *Aethionema* from the Anatolian Diagonal, its cradle of origin (Chapter 3).

Local biotic and abiotic perturbations can also define genomic differences within a species. For example, *Arabidopsis arenosa* (Brassicaceae) has diploid populations, occurring mainly on non-serpentine soil while some of the autotetraploids are adapted to serpentine soils (Arnold *et al.* 2015, Arnold *et al.* 2016). Similarly, I found ploidy differences in our analysis of *Ae. arabicum* (Chapter 4). Our sample sizes were too small to assess whether there is geological, geographical or (a)biotic separation between these populations. Nonetheless the Iranian tetraploid populations already differ in their defence compounds from their diploid counterparts and these differences seem to have been selected for (Chapter 4). I do not yet know whether or what the effect of these differences is on the fitness of the tetraploid populations. My population genetic analyses did confirm that one of Brassicales key-innovations, glucosinolates, plays a role in the genetic diversity of *Ae. arabicum*.

Natural history collections are a valuable resource, not only for phylogenomics and phylogeography (Wandeler *et al.* 2007, Bakker *et al.* 2015, Chapter 3) but also to assess the genetic variation of a species over time. Using markers that have a higher chance in being preserved over time and show allelic variation can help to investigate the change in population dynamics and haplotypes over time (Cozzolino *et al.* 2007, Wandeler *et al.* 2007). Cozzolino *et al.* (2007) used this approach to understand the haplotype loss in the endangered orchid *Anacamptis palustris*. It would be feasible to apply a similar approach to assess the allelic diversity over the distribution of *Ae. arabicum* without the need of field samples. In a broader context: natural history collections should be utilised more often to answer questions on evolutionary change at a genomic level.

## Defend and develop

There is a large body of research that investigates plant development: mainly the molecular background of the transition from vegetative to a generative state (Deng *et al.* 2011, Bouché *et al.* 2016 and the references therein). There is also a great deal of research on the plant defences against herbivores and their fitness effects (reviewed by Hopkins *et al.* 2009). However, these two worlds have rarely come together. An exception to this is the study of Jensen *et al.* (2015, see Chapter 5). We examined changes in glucosinolates throughout the development of two *Ae. arabicum* lines (CYP and TUR). We also identified QTLs that control glucosinolate synthesis within different tissues by using recombinant inbred lines of CYP and TUR (Chapter 5). Fine mapping the major genomic location resulted in two main genes that are involved in the plant development (*SULTR2;1* and *FLC*), but apparently also in the glucosinolate phenotype. This indicates that development and defence correspond at the molecular level, whereby the plant adjusts the levels and composition of its defence compounds depending on its developmental stage, or vice versa. However, as this region consists of a total of 15 genes we have not shown the causality between the molecular background of plant development and plant defence.

Currently there are no successful stable transformation protocols available for *Ae. arabicum*, thus other resources than knock-outs and over-expression analyses need to be used to understand the underlying pathway between development and defence. Moreover, *SULTR2;1* and *FLC* are both involved in basic developmental functions, knock-out experiments would thus have large effects on various facets of plant development. This could make their effect on glucosinolate biosynthesis indiscernible. One way to assess whether *SULTR2;1* and/or *FLC* directly affect glucosinolate content is by fine mapping the region and assessing the glucosinolate phenotype. Gene expression analyses (RT-PCR) of *FLC* and *SULTR2;1* of the fine mapping population might help to investigate whether the phenotypes are due to gene sequences or gene-expression. Experiments involving pathogen and herbivore stress on the (fine mapped) RIL population might help to assess the causality between flowering and glucosinolate composition. These stress-experiments could be combined with gene expression and glucosinolate analyses. RNAi experiments influencing the glucosinolate biosynthesis and measurements of flowering time might also be an option. A high throughput method would be to analyse the epigenome under stressful and control environments to see whether chromatin modifications are affected by changes in flowering time as well as changes in glucosinolate content. Transformation of *Ae. arabicum* alleles into a species that can be transformed (like *A. thaliana*) might help the investigation the molecular pathway between development and defence.

## 6

### The untested model

The Whole Genome Duplication Radiation Lag Time model (WGD-RLT model, Schranz *et al.* 2012) states that diversification lags behind a WGD event. Tank *et al.* (2015) used a diversification model to investigate the presence of a WGD-RLT within the angiosperms. However the WGD-RLT is not the mechanism behind the spurs of diversification seen in the species rich clades (Schranz *et al.* 2012, Tank *et al.* 2015). Key innovations based on geological and climatic effect, but dependent on the polyploidy background, might underlie the success of a lineage (Schranz *et al.* 2012). These innovations do not discern the possibility of extinction events within the species poor clade.

The WGD-RTL encompasses the entire biology of an asymmetrical group. This can vary from the historical biogeography to current genetic diversity within a species and the evolution of traits. These traits are regulated by the coding and non-coding part of the (polyploidy) genome. All these elements influence the life history of an individual. A species life-history encompasses its development, the switch from generative to vegetative life stages and defence traits.

To be able to test the WGD-RTL, we need to know more about the species poor and species rich clades. Here I dived into the species-poor clade of the Brassicaceae, *Aethionema*, trying to understand as much as possible about several different aspects of its biology. One of the challenges for understanding the Brassicaceae species-rich core is the lack of a well-supported phylogeny. However, this is underway (Koch *et al.* personal communication). A well-supported phylogeny of the Brassicaceae family could resolve the questions about the families' centre of origin, their migratory routes and the possibility of extinction events. Especially the latter could resolve the question whether the small size of *Aethionema* is based on the lack of speciation, an increase in extinction an/or is related to a trait. Examples of such studies have been conducted within the Bromelioideae and Solanaceae. In Bromelioideae CAM photosynthesis correlated to net speciation, while the tank habit correlates with extinction (Silvestro *et al.* 2014). Within Solanaceae core group self-incompatibility seems to drive speciation (Goldberg *et al.* 2010).

### **To conclude**

In the last several years we have developed *Aethionema* as a model species for comparative genetics and genomics, but also for ecological important traits (e.g. fruit dimorphism; Lenser *et al.* 2016). In this thesis, I investigated *Aethionema* at its historical, genetic and trait diversity. These themes are required, among others, to test the WGD-RTL model. Moreover it seems as if we have now a natural model that links life history to plant defence, a topic already acknowledged in ecology (Stearns 1989, Van Zandt 2007), but for which its molecular underpinning is still unknown.





## References

- AARABI, F., KUSAJIMA, M., TOHGE, T., KONISHI, T., GIGOLASHVILI, T., TAKAMUNE, M., SASAZAKI, Y., et al.** 2016. Sulfur deficiency-induced repressor proteins optimize glucosinolate biosynthesis in plants. *Science Advances*, **2**, 1–18.
- AGNETA, R., MÖLLERS, C., DE MARIA, S. & RIVELLI, A.R.** 2014. Evaluation of root yield traits and glucosinolate concentration of different *Armoracia rusticana* accessions in Basilicata region (southern Italy). *Scientia Horticulturae*, **170**, 249–255.
- AL-SHEHBAZ, I.A.** 2012. A generic and tribal synopsis of the Brassicaceae (Cruciferae). *Taxon*, **61**, 931–954.
- AL-SHEHBAZ, I.A., BEILSTEIN, M.A. & KELLOGG, E.A.** 2006. Systematics and phylogeny of the Brassicaceae (Cruciferae): an overview. *Plant Systematics and Evolution*, **259**, 89–120.
- ALTSCHUL, S.F., GISH, W., MILLER, W., MYERS, E.W. & LIPMAN, D.J.** 1990. Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403–410.
- AMES, M. & SPOONER, D.M.** 2008. DNA from herbarium specimens settles a controversy about origins of the European potato. *American Journal of Botany*, **95**, 252–257.
- ANSELL, S.W., STENOIEN, H.K., GRUNDMANN, M., RUSSELL, S.J., KOCH, M. A., SCHNEIDER, H. & VOGEL, J.C.** 2011. The importance of Anatolian mountains as the cradle of global diversity in *Arabis alpina*, a key arctic-alpine species. *Annals of Botany*, **108**, 241–252.
- ARNOLD, B., KIM, S., BOMBLIES, K. & BIOLOGY, E.** 2015. Single geographic origin of a widespread autotetraploid. *Molecular Biology and Evolution*, **32**, 1382–1395.
- ARNOLD, B.J., LAHNER, B., DACOSTA, J.M., WEISMAN, C.M., HOLLISTER, J.D., SALT, D.E., BOMBLIES, K. & YANT, L.** 2016. Borrowed alleles and convergence in serpentine adaptation. *Proceedings of the National Academy of Sciences*, **113**, 1–6.
- ARRIGO, N. & BARKER, M.S.** 2012. Rarely successful polyploids and their legacy in plant genomes. *Current Opinion in Plant Biology*, **15**, 140–146.
- AXTELL, M.J.** 2013. Classification and comparison of small RNAs from plants. *Annual Review of Plant Biology*, **64**, 137–159.
- BACHTROG, D. & ANDOLFATTO, P.** 2006. Selection, recombination and demographic history in *Drosophila miranda*. *Genetics*, **174**, 2045–2059.
- BAILEY, C.D., KOCH, M.A., MAYER, M., MUMMENHOFF, K., O'KANE, S.L., WARWICK, S.I., WINDHAM, M.D. & AL-SHEHBAZ, I.A.** 2006. Toward a global phylogeny of the Brassicaceae. *Molecular Biology and Evolution*, **23**, 2142–2160.
- BAKKER, F.T., LEI, D., YU, J., MOHAMMADIN, S., WEI, Z., VAN DE KERKE, S., GRAVENDEEL, B., et al.** 2015. Herbarium genomics: plastome sequence assembly from a range of herbarium specimens using an Iterative Organelle Genome Assembly pipeline. *Biological Journal of the Linnean Society*, **371**, 1–11.
- BANSAL, V. & BAFNA, V.** 2008. HapCUT: An efficient and accurate algorithm for the haplotype assembly problem. *Bioinformatics*, **24**, 153–159.
- BARKER, M.S., VOGEL, H. & SCHRANZ, M.E.** 2009. Paleopolyploidy in the Brassicales: analyses of the *Cleome* transcriptome elucidate the history of genome duplications in *Arabidopsis* and other Brassicales. *Genome Biology and Evolution*, **1**, 391–399.
- BARRETT, C.F., BAKER, W.J., COMER, J.R., CONRAN, J.G., LAHMEYER, S.C., LEEBENS-, J.H., LI, J., et al.** 2016. Plastid genomes reveal support for deep phylogenetic relationships and extensive rate variation among palms and other commelinid monocots. *New Phytologist*, **209**, 855–870.

## References

- BASS, H.W. & BIRCHLER, J.A.** 2012. *Plant Cytogenetics*. Bass, H.W. & Birchler, J.A., eds. New York, NY: Springer New York.
- BATISTA, P.J. & CHANG, H.Y.** 2013. Long noncoding RNAs: Cellular address codes in development and disease. *Cell*, **152**, 1298–1307.
- BEEKWEELDER, J., VAN LEEUWEN, W., VAN DAM, N.M., BERTOSSI, M., GRANDI, V., MIZZI, L., SOLOVIEV, M., et al.** 2008. The impact of the absence of aliphatic glucosinolates on insect herbivory in *Arabidopsis*. *PLoS ONE*, **3**.
- BEILSTEIN, M.A., AL-SHEHBAZ, I.A. & KELLOGG, E.A.** 2006. Brassicaceae phylogeny and trichome evolution. *American Journal of Botany*, **93**, 607–619.
- BEILSTEIN, M.A., AL-SHEHBAZ, I.A., MATHEWS, S. & KELLOGG, E.A.** 2008. Brassicaceae phylogeny inferred from phytochrome A and ndhF sequence data: Tribes and trichomes revisited. *American Journal of Botany*, **95**, 1307–1327.
- BEILSTEIN, M.A., BRINEGAR, A.E. & SHIPPEN, D.E.** 2012. Evolution of the *Arabidopsis* telomerase RNA. *Frontiers in Genetics*, **3**, 1–8.
- BEILSTEIN, M.A., NAGALINGUM, N.S., CLEMENTS, M.D., MANCHESTER, S.R. & MATHEWS, S.** 2010. Dated molecular phylogenies indicate a Miocene origin for *Arabidopsis thaliana*. *Proceedings of the National Academy of Sciences*, **107**, 18724–18728.
- BERGH, E. VAN DEN, HOFBERGER, J.A. & SCHRANZ, M.E.** 2016. Flower power and the mustard bomb: Comparative analysis of gene and genome duplications in glucosinolate biosynthetic pathway evolution in Cleomaceae and Brassicaceae. *American Journal of Botany*, **103**, 1212–1222.
- BIBALANI, G.H.** 2012. Investigation on flowering phenology of Brassicaceae in the Shanjan region Shabestar district, NW Iran (usage for honeybees. *Annals of Biological Research*, **6**, 1958–1968.
- BLANC, G. & WOLFE, K.H.** 2004. Functional divergence of duplicated genes formed by polyploidy during *Arabidopsis* evolution. *The Plant Cell*, **16**, 1679–1691.
- BOERNER, S. & MCGINNIS, K.M.** 2012. Computational identification and functional predictions of long noncoding RNA in *Zea mays*. *PLoS ONE*, **7**, e43047.
- BOLGER, A.M., LOHSE, M. & USADEL, B.** 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.
- BOUCHÉ, F., WOODS, D. & AMASINO, R.M.** 2016. Winter memory throughout the plant kingdom: different paths to flowering. *Plant Physiology*, pp.01322.2016.
- BOWERS, J.E., CHAPMAN, B.A. & RONG, J.** 2003. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature*, **422**, 433–438.
- BROCHMANN, C.** 1993. Reproductive strategies of diploid and polyploid populations of arctic *Draba* (Brassicaceae). *Plant Systematics and Evolution*, **185**, 55–83.
- BROMAN, K.W., WU, H., SEN, S. & CHURCHILL, G.A.** 2003. R/qtl: QTL mapping in experimental crosses. *Bioinformatics*, **19**, 889–890.
- BROWN, P.D., TOKUHISA, J.G., REICHEL, M. & GERSHENZON, J.** 2003. Variation of glucosinolate accumulation among different organs and developmental stages of *Arabidopsis thaliana*. *Phytochemistry*, **62**, 471–481.
- BUROW, M., MÜLLER, R., GERSHENZON, J. & WITTSTOCK, U.** 2006. Altered glucosinolate hydrolysis in genetically engineered *Arabidopsis thaliana* and its influence on the larval development of *Spodoptera littoralis*. *Journal of Chemical Ecology*, **32**, 2333–2349.
- CAMACHO, C., COULOURIS, G., AVAGYAN, V., MA, N., PAPADOPOULOS, J., BEALER, K. & MADDEN, T.L.** 2009. BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.
- CARDINAL-McTEAGUE, W.M., SYTSMA, K.J. & HALL, J.C.** 2016. Biogeography and diversification of Brassicales : A 103 million year tale. *Molecular Phylogenetics and Evolution*, **99**, 204–224.

- CHENG, S., VAN DEN BERGH, E., ZENG, P., ZHONG, X., XU, J., LIU, X., HOFBERGER, J., et al.** 2013. The *Tarenaya hassleriana* genome provides insight into reproductive trait and genome evolution of crucifers. *The Plant Cell*, **25**, 2813–2830.
- COLEMAN, F.O.D., BLAKE-KALFF, M.M.A. & DAVIES, T.E.** 1997. Detoxification of xenobiotics by plants: chemical modification and vacuolar compartmentation. *Trends in Plant Science*, **2**, 144–151.
- COMAI, L.** 2006. The advantages and disadvantages of being introduced. *Nature Reviews Genetics*, **6**, 836–846.
- COMPUTING, R. FOUNDATION FOR S.** 2013. R: A language and environment for statistical computing.
- COUVREUR, T.L.P., FRANZKE, A., AL-SHEHBAZ, I.A., BAKKER, F.T., KOCH, M. A & MUMMENHOFF, K.** 2010. Molecular phylogenetics, temporal diversification, and principles of evolution in the mustard family (Brassicaceae). *Molecular Biology and Evolution*, **27**, 55–71.
- COZZOLINO, S., CAFASSO, D., PELLEGRINO, G., MUSACCHIO, A. & WIDMER, A.** 2007. Genetic variation in time and space: The use of herbarium specimens to reconstruct patterns of genetic variation in the endangered orchid *Anacamptis palustris*. *Conservation Genetics*, **8**, 629–639.
- CSORBA, T., QUESTA, J.I., SUN, Q. & DEAN, C.** 2014. Antisense COOLAIR mediates the coordinated switching of chromatin states at FLC during vernalization. *Proceedings of the National Academy of Sciences*, **111**, 16160–16165.
- DANECEK, P., AUTON, A., ABECASIS, G., ALBERS, C.A., BANKS, E., DEPRISTO, M.A., HANDSAKER, R.E., et al.** 2011. The variant call format and VCFtools. *Bioinformatics*, **27**, 2156–2158.
- DAVIS, P.H.** 1965. Cruciferae. In: Davis, P.H., ed. *Flora of Turkey*. Edinburgh, UK: Edinburgh University Press, 321–330.
- DAVIS, P.H.** 1971. Distribution patterns in Anatolia with particular reference to endemism. In: Davis, P.H., Harper, P.C. & Hedge, I.C., eds. *Plant Life of South-West Asia*. Royal Botanic Garden, 15–28.
- DE SMET, R., ADAMS, K.L., VANDEPOELE, K., VAN MONTAGU, M.C.E., MAERE, S. & VAN DE PEER, Y.** 2013. Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. *Proceedings of the National Academy of Sciences*, **110**, 2898–2903.
- DENG, W., YING, H., HELLIWELL, C.A., TAYLOR, J.M., PEACOCK, W.J. & DENNIS, E.S.** 2011. FLOWERING LOCUS C (FLC) regulates development pathways throughout the life cycle of *Arabidopsis*. *Proceedings of the National Academy of Sciences*, **108**, 6680–6685.
- DEPRISTO, M.A., BANKS, E., POPLIN, R., GARIMELLA, K. V, MAGUIRE, J.R., HARTL, C., PHILIPPAKIS, A.A., et al.** 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, **43**, 491–498.
- DI, C., YUAN, J., WU, Y., LI, J., LIN, H., HU, L., ZHANG, T., et al.** 2014. Characterization of stress-responsive lncRNAs in *Arabidopsis thaliana* by integrating expression, epigenetic and structural features. *The Plant Journal*, **80**, 848–861.
- DIERSCHKE, T., MANDÁKOVÁ, T., LYSÁK, M. A. & MUMMENHOFF, K.** 2009. A bicontinental origin of polyploid Australian/New Zealand *Lepidium* species (Brassicaceae): evidence from genomic in situ hybridization (GISH). *Annals of Botany*, **104**, 681–688.
- DING, J., LU, Q., OUYANG, Y., MAO, H., ZHANG, P., YAO, J., XU, C. & LI, X.** 2012a. A long noncoding RNA regulates photoperiod-sensitive male sterility, an essential component of hybrid rice. *Proceedings of the National Academy of Sciences*, **109**, 2654–2659.

## References

- DING, J., SHEN, J., MAO, H., XIE, W., LI, X. & ZHANG, Q.** 2012b. RNA-directed DNA methylation is involved in regulating photoperiod-sensitive male sterility in rice. *Molecular plant*, **5**, 1210–1216.
- DINGER, M.E., PANG, K.C., MERCER, T.R. & MATTICK, J.S.** 2008. Differentiating protein-coding and noncoding RNA: Challenges and ambiguities. *PLoS Computational Biology*, **4**.
- DJAMALI, M., BAUMEL, A., BREWER, S., JACKSON, S.T., KADEREIT, J.W., LOPEZ-VINYALLONGA, S., MEHREGAN, I., SHABANIAN, E. & SIMAKOVA, A.** 2012. Ecological implications of *Cousinia* Cass. (Asteraceae) persistence through the last two glacial-interglacial cycles in the continental Middle East for the Irano-Turanian flora. *Review of Palaeobotany and Palynology*, **172**, 10–20.
- DOLATYARI, A., VALLÈS, J., NAGHAVI, M.R. & SHAHZADEH FAZELI, S.A.** 2013. Karyological data of 47 accessions of 28 *Artemisia* (Asteraceae, Anthemideae) species from Iran, with first new reports for Iranian populations and first absolute counts in three species. *Plant Systematics and Evolution*, **299**, 1503–1518.
- DOLEZELT, J., GREILHUBER, J., LUCRETTIII, S., MEISTER, A., LYSAKT, M.A. & NARDIII, L.** 1998. Plant genome size estimation by flow cytometry : Inter-laboratory comparison. *Annals of Botany*, **82**, 17–26.
- DRUMMOND, A.J., HO, S.Y.W., PHILLIPS, M.J. & RAMBAUT, A.** 2006. Relaxed phylogenetics and dating with confidence. *PLoS Biology*, **4**, 699–710.
- DRUMMOND, A.J., SUCHARD, M.A., XIE, D. & RAMBAUT, A.** 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution*, **29**, 1969–1973.
- DUARTE, J.M., WALL, P.K., EDGER, P.P., LANDHERR, L.L., MA, H., PIRES, J.C., LEEBENS-MACK, J. & CLAUDE, W.** 2010. Identification of shared single copy nuclear genes in *Arabidopsis*, *Populus*, *Vitis* and *Oryza* and their phylogenetic utility across various taxonomic levels. *BMC Evolutionary Biology*, **10**, 1–18.
- DUVICK, J., FU, A., MUPPIRALA, U., SABHARWAL, M., WILKERSON, M.D., LAWRENCE, C.J., LUSHBOUGH, C. & BRENDDEL, V.** 2008. PlantGDB: a resource for comparative plant genomics. *Nucleic Acids Research*, **36**, D959–D965.
- EDGER, P.P., HEIDEL-FISCHER, H.M., BEKAERT, M., ROTA, J., GLÖCKNER, G., PLATTS, A.E., HECKEL, D.G., et al.** 2015. The butterfly plant arms-race escalated by gene and genome duplications. *Proceedings of the National Academy of Sciences*, **112**, 8362–8366.
- ERTUGRUL, K. & BEYAZGLU, O.** 1996. A new species from south Anatolia - *Aethionema karamanicus* (Cruciferae). *Turkish Journal of Botany*, **21**, 99–101.
- FALK, K.L., TOKUHISA, J.G. & GERSHENZON, J.** 2007. The effect of sulfur nutrition on plant glucosinolate content: Physiology and molecular mechanisms. *Plant Biology*, **9**, 573–581.
- FAWCETT, J.A., MAERE, S. & PEER, Y. VAN DE.** 2009. Plants with double genomes might have had a better chance to survive the Cretaceous– Tertiary extinction event. *Proceedings of the National Academy of Sciences*, **106**, 5737–5742.
- FEDAK, H., PALUSINSKA, M., KRZYCZMONIK, K., BRZEZNIAK, L. & YATUSEVICH, R.** 2016. Control of seed dormancy in *Arabidopsis* by a cis-acting noncoding antisense transcript. *Proceedings of the National Academy of Sciences*, **113**, E7846–E7855.
- FLINTOFT, L.** 2013. Non-coding RNA: Structure and function for lncRNAs. *Nature Reviews Genetics*, **14**, 598.
- FRANZKE, A., GERMAN, D., AL-SHEHBAB, I.A. & MUMMENHOFF, K.** 2009. *Arabidopsis* family ties: molecular phylogeny and age estimates in Brassicaceae. *Taxon*, **58**, 425–437.
- FRANZKE, A., KOCH, M.A. & MUMMENHOFF, K.** 2016. Turnip time travels: age estimates in Brassicaceae. *Trends in Plant Science*, **21**, 554–561.

- FRANZKE, A., LYSAK, M.A., AL-SHEHBAZ, I.A., KOCH, M.A. & MUMMENHOFF, K.** 2011. Cabbage family affairs: the evolutionary history of Brassicaceae. *Trends in Plant Science*, **16**, 108–116.
- FREELING, M., LYONS, E., PEDERSEN, B., ALAM, M., MING, R. & LISCH, D.** 2008. Many or most genes in *Arabidopsis* transposed after the origin of the order Brassicales. *Genome Research*, **18**, 1924–1937.
- FRIDMAN, E. & ZAMIR, D.** 2003. Functional divergence of a syntenic invertase gene family in tomato, potato, and *Arabidopsis*. *Plant physiology*, **131**, 603–609.
- FROVA, C.** 2006. Glutathione transferases in the genomics era: New insights and perspectives. *Biomolecular Engineering*, **23**, 149–169.
- GBIF.** 2012. Recommended practices for citation of data published through the GBIF network (Authored By Vishwas Chavan). *Copenhagen Global Biodiversity Information Facility*, 1–12.
- GIGOLASHVILI, T. & KOPRIVA, S.** 2014. Transporters in plant sulfur metabolism. *Frontiers in Plant Science*, **5**, 442.
- GIVNISH, T.J.** 2010. Ecology of plant speciation. *Taxon*, **59**, 1326–1366.
- GOLDBERG, E.E., KOHN, J.R., LANDE, R., ROBOTSON, K.A., SMITH, S.A. & IGIC, B.** 2010. Species selection maintains self-incompatibility. *Science*, **330**, 493–495.
- GOODSTEIN, D.M., SHU, S., HOWSON, R., NEUPANE, R., HAYES, R.D., FAZO, J., MITROS, T., et al.** 2012. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Research*, **40**, D1178–86.
- GOODWIN, Z.A., HARRIS, D.J., FILER, D., WOOD, J.R.I. & SCOTLAND, R.W.** 2015. Widespread mistaken identity in tropical plant collections. *Current Biology*, **25**, R1066–R1067.
- GOSSMANN, T.I., SONG, B.-H., WINDSOR, A.J., MITCHELL-OLDS, T., DIXON, C.J., KAPRALOV, M. V., FILATOV, D.A. & EYRE-WALKER, A.** 2010. Genome wide analyses reveal little evidence for adaptive evolution in many plant species. *Molecular Biology and Evolution*, **27**, 1822–1832.
- GRABHERR, M.G., HAAS, B.J., YASSOUR, M., LEVIN, J.Z., THOMPSON, J.D., AMIT, I., ADICONIS, X., et al.** 2011. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nature Biotechnology*, **29**, 644–652.
- GRIFFITHS-JONES, S., SAINI, H.K., VAN DONGEN, S. & ENRIGHT, A.J.** 2008. miRBase: tools for microRNA genomics. *Nucleic Acids Research*, **36**, D154–8.
- GRUBER, A.R., LORENZ, R., BERNHART, S.H., NEUBÖCK, R. & HOFACKER, I.L.** 2008. The Vienna RNA websuite. *Nucleic Acids Research*, **36**, W70–4.
- GÜL, S.** 2013. Ecological divergence between two evolutionary lineages of *Hyla savignyi* (Audouin, 1827) in Turkey: Effects of the Anatolian Diagonal. *Animal Biology*, **63**, 285–295.
- GUTTMAN, M., AMIT, I., GARBER, M., FRENCH, C., LIN, M.F., HUARTE, M., ZUK, O., et al.** 2009. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature*, **458**, 223–227.
- HA, M., KIM, E. & CHEN, Z.J.** 2009. Duplicate genes increase expression diversity in closely related species and allopolyploids. *Proceedings of the National Academy of Sciences*, **106**, 2295–2300.
- HALKIER, B.A. & GERSHENZON, J.** 2006. Biology and biochemistry of glucosinolates. *Annual Review of Plant Biology*, **57**, 303–333.
- HANADA, K., ZHANG, X., BOREVITZ, J.O., LI, W.-H. & SHIU, S.-H.** 2007. A large number of novel coding small open reading frames in the intergenic regions of the *Arabidopsis thaliana* genome are transcribed and/or under purifying selection. *Genome Research*, **17**, 632–640.
- HAUDRY, A., PLATTS, A.E., VELLO, E., HOEN, D.R., LECLERCQ, M., WILLIAMSON, R.J., FORCZEK, E., et al.** 2013. An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. *Nature Genetics*, **45**, 891–898.

## References

- HAYEK, A.** 1911. Entwurf eines Cruciferensystems auf phylogenetischer Grundlage. *Beihefte des Botanischen Centralblattes*, **27**, 127–335.
- HAYES, J.D., FLANAGAN, J.U. & JOWSEY, I.R.** 2005. Glutathione transferases. *Annual Review of Pharmacology*, **45**, 51–88.
- HEDGE, I.C.** 1976. A systematic and geographical survey of the old world cruciferae. In Vaughan, J.G., Macleod, A.J. & Jones, B.M., eds. *The Biology and Chemistry of the Cruciferae*. London, UK: Academic Press, 1–35.
- HEDGE, I.C.** 1965. *Aethionema*. In Davis, P.H., Cullen, I. & Coode, M.J., eds. *Flora of Turkey and the East Aegean Islands*. Edinburgh, UK: Edinburgh University Press, 314–330.
- HOFBERGER, J.A., LYONS, E., EDGER, P.P., CHRIS PIRES, J. & ERIC SCHRANZ, M.** 2013. Whole genome and tandem duplicate retention facilitated glucosinolate pathway diversification in the mustard family. *Genome Biology and Evolution*, **5**, 2155–2173.
- HOHMANN, N., WOLF, E.M., LYSAK, M. A. & KOCH, M. A.** 2015. A time-calibrated road map of Brassicaceae species radiation and evolutionary history. *The Plant Cell*, **27**, 2770–2784.
- HOPKINS, R.J., VAN DAM, N.M. & VAN LOON, J.J.A.** 2009. Role of glucosinolates in insect-plant relationships and multitrophic interactions. *Annu. Rev. Entomol*, **54**, 57–83.
- HUALA, E., DICKERMAN, A.W., GARCIA-HERNANDEZ, M., WEEMS, D., REISER, L., LAFOND, F., HANLEY, D., et al.** 2001. The *Arabidopsis* Information Resource (TAIR): a comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant. *Nucleic Acids Research*, **29**, 102–105.
- HUANG, C., SUN, R., HU, Y., ZENG, L., ZHANG, N., CAI, L., ZHANG, Q., et al.** 2016. Resolution of Brassicaceae phylogeny using nuclear genes uncovers nested radiations and supports convergent morphological evolution. *Molecular Biology and Evolution*, **33**, 394–412.
- HUELSENBECK, J.P. & RONQUIST, F.** 2001. MrBayes: Bayesian inference of phylogenetic trees. *Bioinformatics*, **17**, 754–755.
- HUSON, D.H. & BRYANT, D.** 2006. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution*, **23**, 254–267.
- IETSWAART, R., WU, Z. & DEAN, C.** 2012. Flowering time control: Another window to the connection between antisense RNA and chromatin. *Trends in Genetics*, **28**, 445–453.
- IMBERT, E.** 2002. Ecological consequences and ontogeny of seed heteromorphism. *Perspectives in Plant Ecology Evolution and Systematics*, **5**, 13–36.
- JANCHEN, E.** 1942. Das System der Cruciferen. *Österreichische Botanische Zeitschrift*, **91**, 1–28.
- JENSEN, L.M., JEPSEN, H.S.K., HALKIER, B.A., KLIEBENSTEIN, D.J. & BUROW, M.** 2015. Natural variation in cross-talk between glucosinolates and onset of flowering in *Arabidopsis*. *Frontiers in Plant Science*, **6**, 697.
- JIAO, Y., WICKETT, N.J., AYYAMPALAYAM, S., CHANDERBALI, A.S., LANDHERR, L., RALPH, P.E., TOMSHO, L.P., et al.** 2011. Ancestral polyploidy in seed plants and angiosperms. *Nature*, **473**, 97–100.
- JIN, J., LIU, J., WANG, H., WONG, L. & CHUA, N.-H.** 2013. PLncDB: plant long non-coding RNA database. *Bioinformatics*, **29**, 1068–1071.
- JORDON-THADEN, I.E.** 2009. Species and genetic diversity of *Draba*: phylogeny and phylogeography. PhD Dissertation. *Faculty of Biological Sciences, University of Heidelberg, Germany*.
- JORDON-THADEN, I.E., AL-SHEHBAZ, I.A. & KOCH, M.A.** 2013. Species richness of the globally distributed, arctic-alpine genus *Draba* L. (Brassicaceae). *Alpine Botany*, **123**, 97–106.
- KAGALE, S., ROBINSON, S.J., NIXON, J., XIAO, R., HUEBERT, T., CONDIE, J., KESSLER, D., et al.** 2014. Polyploid evolution of the Brassicaceae during the Cenozoic Era. *The Plant Cell*, **26**, 2777–2791.

- KANE, M.J., EMERSON, J.W. & WESTON, S.** 2013. Scalable strategies for computing with massive data. *JSS Journal of Statistical Software*, **55**.
- KARL, R. & KOCH, M.A.** 2013. A world-wide perspective on crucifer speciation and evolution: Phylogenetics, biogeography and trait evolution in tribe Arabideae. *Annals of Botany*, **112**, 983–1001.
- KATOH, K. & STANDLEY, D.M.** 2013. MAFFT multiple sequence alignment software version7: improvements in performance and usability. *Molecular Biology and Evolution*, **30**, 772–780.
- KEARSE, M., MOIR, R., WILSON, A., STONES-HAVAS, S., CHEUNG, M., STURROCK, S., BUXTON, S., et al.** 2012. Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, **28**, 1647–1649.
- KEELER, K.H.** 1992. Local polyploid variation in the native prairie grass *Andropogon gerardii*. *American Journal of Botany*, **79**, 1229–1232.
- KLIEBENSTEIN, D.J.** 2004. Secondary metabolites and plant/environment interactions: A view through *Arabidopsis thaliana* tinted glasses. *Plant, Cell and Environment*, **27**, 675–684.
- KLIEBENSTEIN, D.J., GERSHENZON, J. & MITCHELL-OLDS, T.** 2001. Comparative quantitative trait loci mapping of aliphatic, indolic and benzylic glucosinolate production in *Arabidopsis thaliana* leaves and seeds. *Genetics*, **159**, 359–370.
- KLIEBENSTEIN, D.J., KROYMANN, J., BROWN, P., FIGUTH, A., PEDERSEN, D., GERSHENZON, J. & MITCHELL-OLDS, T.** 2001. Genetic control of natural variation in *Arabidopsis* glucosinolate accumulation. *Plant Physiology*, **126**, 811–825.
- KNILL, T., SCHUSTER, J., REICHEL, M., GERSHENZON, J. & BINDER, S.** 2008. *Arabidopsis* branched-chain aminotransferase 3 functions in both amino acid and glucosinolate biosynthesis. *Plant Physiology*, **146**, 1028–1039.
- KOCH, M.A. & KIEFER, C.** 2006. Molecules and migration: Biogeographical studies in cruciferous plants. *Plant Systematics and Evolution*, **259**, 121–142.
- KOENIG, D. & WEIGEL, D.** 2015. Beyond the thale: comparative genomics and genetics of *Arabidopsis* relatives. *Nature Reviews Genetics*, **16**, 285–298.
- KOMAROV, V.L.** 1934. Flora URSS (Flora Unionis Rerumpublicarum Sovieticarum Socialistarum).
- KOROLEVA, O.A., DAVIES, A., DEEKEN, R., THORPE, M.R., TOMOS, A.D. & HEDRICH, R.** 2000. Different myrosinase and ideoblast distribution in *Arabidopsis* and *Brassica napus*. *Plant Physiology*, **127**, 1750–1763.
- LANFEAR, R., CALCOTT, B., HO, S.Y.W. & GUINDON, S.** 2012. PartitionFinder: Combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Molecular Biology and Evolution*, **29**, 1695–1701.
- LENSER, T., GRAEBER, K., CEVIK, Ö.S., ADIGÜZEL, N., DÖNMEZ, A.A., KETTERMANN, M., MAYLAND-QUELLHORST, S., et al.** 2016. *Aethionema arabicum* as a model system for studying developmental control and plasticity of fruit and seed dimorphism. *Plant physiology*, **172**, 1691–1707.
- LEVY-RIMLER, G., BELL, R.E., BEN-TAL, N. & AZEM, A.** 2002. Type I chaperonins: not all are created equal. *FEBS letters*, **529**, 1–5.
- LI, L., EICHTEN, S.R., SHIMIZU, R., PETSCH, K., YEH, C.-T., WU, W., CHETTOOR, A.M., et al.** 2014. Genome-wide discovery and characterization of maize long non-coding RNAs. *Genome Biology*, **15**, R40.
- LI, L., JR, C.J.S. & ROOS, D.S.** 2003. OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Resource*, **13**, 2178–2189.

## References

- LI, L., WANG, X., SASIDHARAN, R., STOLC, V., DENG, W., HE, H., KORBEL, J., et al. 2007. Global identification and characterization of transcriptionally active regions in the rice genome. *PLoS ONE*, **2**, e294.
- LIU, J., JUNG, C., XU, J., WANG, H., DENG, S., BERNAD, L., ARENAS-HUERTERO, C. & CHUA, N.-H. 2012. Genome-wide analysis uncovers regulation of long intergenic noncoding RNAs in *Arabidopsis*. *The Plant Cell*, **24**, 4333–4345.
- LUIKART, G., ENGLAND, P.R., TALLMON, D., JORDAN, S. & TABERLET, P. 2003. The power and promise of population genomics: from genotyping to genome typing. *Nature reviews. Genetics*, **4**, 981–994.
- LYONS, E. & FREELING, M. 2008. How to usefully compare homologous plant genes and chromosomes as DNA sequences. *The Plant Journal*, **53**, 661–673.
- LYSAK, M.A., MANDÁKOVÁ, T. & SCHRANZ, M.E. 2016. Comparative paleogenomics of crucifers: ancestral genomic blocks revisited. *Current Opinion in Plant Biology*, **30**, 108–115, 10.1016/j.pbi.2016.02.001.
- MA, Z., CORUH, C. & AXTELL, M.J. 2010. *Arabidopsis lyrata* small RNAs: transient MIRNA and small interfering RNA loci within the *Arabidopsis* genus. *The Plant Cell*, **22**, 1090–1103.
- MACAS, J., MESZAROS, T. & NOUZOVA, M. 2002. PlantSat: a specialized database for plant satellite repeats. *Bioinformatics*, **18**, 28–35.
- MALIK, M.S., RILEY, M.B., NORSWORTHY, J.K. & BRIDGES, W. 2010. Variation of glucosinolates in wild radish (*Raphanus raphanistrum*) accessions. *Journal of Agricultural and Food Chemistry*, **58**, 11626–11632.
- MANAFZADEH, S., STAEDLER, Y.M. & CONTI, E. 2016. Visions of the past and dreams of the future in the Orient: the Irano-Turanian region from classical botany to evolutionary studies. *Biological Reviews*, **9**, 1–24.
- MANO, Y. & NEMOTO, K. 2012. The pathway of auxin biosynthesis in plants. *Journal of Experimental Botany*, **63**, 2853–2872.
- MARHOLD, K., KUDOH, H., PAK, J.H., WATANABE, K., Španiel, S. & LIHOVÁ, J. 2010. Cytotype diversity and genome size variation in eastern Asian polyploid *Cardamine* (Brassicaceae) species. *Annals of Botany*, **105**, 249–264.
- MARRS, K.A. 1996. The functions and regulation of glutathione S-transferases in plants. *Annual Review of Plant Physiology and Plant Molecular Biology*, **47**, 127–158.
- MATTICK, J.S. & GAGEN, M.J. 2001. The evolution of controlled multitasked gene networks: The role of introns and other noncoding RNAs in the development of complex organisms. *Molecular Biology and Evolution*, **18**, 1611–1630.
- MAYROSE, I., ZHAN, S.H., ROTHFELS, C.J., MAGNUSON-FORD, K., BARKER, M.S., RIESEBERG, L.H. & OTTO, S.P. 2011. Recently formed polyploid plants diversify at lower rates. *Science*, **333**, 1257.
- MCKENNA, A., HANNA, M., BANKS, E., SIVACHENKO, A.Y., CIBULSKIS, K., KERNYTSKY, A.M., GARIMELLA, K. V, et al. 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, **20**, 1297–1303.
- MÉDAIL, F. & DIADEMA, K. 2009. Glacial refugia influence plant diversity patterns in the Mediterranean Basin. *Journal of Biogeography*, **36**, 1333–1345.
- MICÓ, E., SANMARTÍN, I. & GALANTE, E. 2009. Mediterranean diversification of the grass-feeding *Anisopliina* beetles (Scarabaeidae, Rutelinae, *Anomalini*) as inferred by bootstrap-averaged dispersal-vicariance analysis. *Journal of Biogeography*, **36**, 546–560.
- MILLER, M.A., PFEIFFER, W. & SCHWARTZ, T. 2010. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. *2010 Gateway Computing Environments Workshop, GCE 2010*.



- MOHAMMADIN, S., EDGER, P.P., PIRES, J.C. & SCHRANZ, M.E.** 2015. Positionally-conserved but sequence-diverged: Identification of long non-coding RNAs in the Brassicaceae and Cleomaceae. *BMC Plant Biology*, **15**, 217.
- MOLDOVAN, D., SPRIGGS, A., YANG, J., POGSON, B.J., DENNIS, E.S. & WILSON, I.W.** 2010. Hypoxia-responsive microRNAs and trans-acting small interfering RNAs in *Arabidopsis*. *Journal of Experimental Botany*, **61**, 165–177.
- MÜHLHAUSEN, A., LENSER, T., MUMMENHOFF, K. & THEISSEN, G.** 2013. Evidence that an evolutionary transition from dehiscent to indehiscent fruits in *Lepidium* (Brassicaceae) was caused by a change in the control of valve margin identity genes. *Plant Journal*, **73**, 824–835.
- MUMMENHOFF, K., BRÜGGEMANN, H. & BOWMAN, J.L.** 2001. Chloroplast DNA phylogeny and biogeography of *Lepidium* (Brassicaceae). *American Journal of Botany*, **88**, 2051–2063.
- NOVIKOVA, I. V., HENNELLY, S.P., SANBONMATSU, K.Y. & RNA, K.** 2012. Sizing up long non-coding RNAs: Do lncRNAs have secondary and tertiary structure. *BioArchitecture*, **2**, 189–199.
- OHNO, S.** 1970. Evolution by gene duplication. In *Evolution by gene duplication*. New York, NY: Springer Verlag.
- OLSON-MANNING, C.F., STROCK, C.F. & MITCHELL-OLDS, T.** 2015. Flux control in a defense pathway in *Arabidopsis thaliana* is robust to environmental perturbations and controls variation in adaptive traits. *G3*, **5**, 2421–2427.
- OLSON, D.M., DINERSTEIN, E., WIKRAMANAYAKE, E.D., BURGESS, N.D., POWELL, G.V.N., UNDERWOOD, E.C., D'AMICO, J. A., et al.** 2001. Terrestrial ecoregions of the world: A new map of life on earth. *BioScience*, **51**, 933.
- Özüdoğru, B., AKAYDIN, G., ERIK, S., AL-SHEHBAZ, I.A. & MUMMENHOFF, K.** 2015. Phylogeny, diversification and biogeographic implications of the eastern Mediterranean endemic genus *Ricotia* (Brassicaceae). *Taxon*, **64**, 727–740.
- PALAZZO, A.F. & GREGORY, T.R.** 2014. The Case for Junk DNA. *PLoS Genetics*, **10**, e1004351.
- PANG, K.C., FRITH, M.C. & MATTICK, J.S.** 2006. Rapid evolution of noncoding RNAs: Lack of conservation does not mean lack of function. *Trends in Genetics*, **22**, 1–5.
- PARADIS, E., CLAUDE, J. & STRIMMER, K.** 2004. APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics*, **20**, 289–290.
- PAROLLY, G., NORDT, B., BLEEKER, W. & MUMMENHOFF, K.** 2010. *Heldreichia* Boiss. (Brassicaceae) revisited: A morphological and molecular study. *Taxon*, **59**, 187–202.
- PAVLOVA, D.** 2007. A new species of *Aethionema* (Brassicaceae) from the Bulgarian flora. *Botanical Journal of the Linnean Society*, **155**, 533–540.
- PAYNE, R.W., MURRAY, D.A., HARDING, S.A., BAIRD, D.B. & SOUTAR, D.M.** 2009. GenStat for Windows (12th Edition) Introduction.
- PETERSEN, B.L., CHEN, S., HANSEN, C.H., OLSEN, C.E. & HALKIER, B.A.** 2002. Composition and content of glucosinolates in developing *Arabidopsis thaliana*. *Planta*, **214**, 562–571.
- PRANTL, K.** 1891. Cruciferae. In Engler, A. & Prantl, K., eds. *Die natürlichen Pflanzenfamilien IIIb*. Leipzig: Verlag von Wilhelm Engelmann, 145–206.
- PRICE, R.A., PALMER, J.D. & AL-SHEHBAZ, I.A.** 1994. Systematic relationships of *Arabidopsis*: a molecular and morphological perspective. In M. Meyerowitz & C. R. Somerville (eds.), ed. *Arabidopsis*. New York, NY: Cold Springs Harbor Laboratory Press, 7–19.
- PRITCHARD, J.K., STEPHENS, M. & DONNELLY, P.** 2000. Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.
- RAMBAUT, A., SUCHARD, M., XIE, D. & DRUMMOND, A.** 2014. Tracer v1.6. <http://beast.bio.ed.ac.uk/Tracer>.

## References

- RECHINGER, K.H.** 1968. Cruciferae. In: *Flora Iranica: Flora des iranischen Hochlandes und der umrahmenden Gebirge*. Graz, Austria: Akademische Druck u. Verlagsanstalt, 1–6.
- REDOVNIKOVIC, I.R., GLIVETIC, T., DELONGA, K. & VORKAPIC-FURAC, J.** 2008. Glucosinolates and their potential role in plant. *Periodicum Biologorum*, **110**, 297–309.
- REEVES, R.D. & ADIGÜZEL, N.** 2008. The nickel hyperaccumulating plants of the serpentines of Turkey and adjacent areas: A review with new data. *Turkish Journal of Biology*, **32**, 143–153.
- ROHR, F., ULRICHS, C. & MEWIS, I.** 2009. Variability of aliphatic glucosinolates in *Arabidopsis thaliana* (L.)-Impact on glucosinolate profile and insect resistance. *Journal of Applied Botany and Food Quality*, **82**, 131–135.
- RONQUIST, F. & HUENSENBECK, J.P.** 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, **19**, 1572–1574.
- RYU, H.Y., KIM, S.Y., PARK, H.M., YOU, J.Y., KIM, B.H., LEE, J.S. & NAM, K.H.** 2009. Modulations of AtGSTF10 expression induce stress tolerance and BAK1-mediated cell death. *Biochemical and Biophysical Research Communications*, **379**, 417–422.
- SCHMICKL, R., PAULE, J., KLEIN, J., MARHOLD, K. & KOCH, M.A.** 2012. The evolutionary history of the *Arabidopsis arenosa* complex: Diverse tetraploids mask the Western Carpathian center of species and genetic diversity. *PLoS ONE*, **7**, e42691.
- SCHMUTZ, J., CANNON, S.B., SCHLUETER, J., MA, J., MITROS, T., NELSON, W., HYTEN, D.L., et al.** 2010. Genome sequence of the palaeopolyploid soybean. *Nature*, **463**, 178–183.
- SCHRANZ, M.E., DOBEŠ, C., KOCH, M.A. & MITCHELL-OLDS, T.** 2005. Sexual reproduction, hybridization, apomixis, and polyploidization in the genus *Boechera* (Brassicaceae). *American Journal of Botany*, **92**, 1797–1810.
- SCHRANZ, M.E., MOHAMMADIN, S. & EDGER, P.P.** 2012. Ancient whole genome duplications, novelty and diversification: the WGD Radiation Lag-Time Model. *Current Opinion in Plant Biology*, **15**, 147–153.
- SCHULZ, O.E.** 1936. Cruciferae. In Engler, A. & Prantl, K., eds. *Die natürlichen Pflanzenfamilien*, vol. 17B. Leipzig, Germany: Verlag Wilhelm Engelmann, 227–658.
- SEBASTIAN, P., SCHAEFER, H. & RENNER, S.S.** 2010. Darwin's Galapagos gourd: Providing new insights 175 years after his visit. *Journal of Biogeography*, **37**, 975–978.
- SHARBEL, T.F., MITCHELL-OLDS, T. & DOBEŠ, C.** 2005. Biogeographic distribution of polyploidy and B chromosomes in the apomictic *Boechera holboellii* complex. *Cytogenetic and Genome Research*, **292**, 283–292.
- SILVESTRO, D., ZIZKA, G. & SCHULTE, K.** 2014. Disentangling the effects of key innovations on the diversification of Bromelioideae (Bromeliaceae). *Evolution*, **68**, 163–175.
- SLOTTE, T., FOXE, J.P., HAZZOURI, K.M. & WRIGHT, S.I.** 2010. Genome-wide evidence for efficient positive and purifying selection in *Capsella grandiflora*, a plant species with a large effective population size. *Molecular Biology and Evolution*, **27**, 1813–1821.
- SLOTTE, T., HUANG, H., LASCoux, M. & CEPLITIS, A.** 2008. Polyploid speciation did not confer instant reproductive isolation in *Capsella* (Brassicaceae). *Molecular Biology and Evolution*, **25**, 1472–1481.
- SOLTIS, D.E., ALBERT, V.A., LEEBENS-MACK, J., BELL, C.D., PATERSON, A.H., ZHENG, C., SANKOFF, D., DEPAMPHILIS, C.W., WALL, P.K. & SOLTIS, P.S.** 2009. Polyploidy and angiosperm diversification. *American Journal of Botany*, **96**, 336–348.
- SOLTIS, D.E., BUGGS, R.J.A., DOYLE, JEFF, J. & SOLTIS, P.S.** 2010. What we still don't know about polyploidy. *Taxon*, **59**, 1387–1403.
- SØNDERBY, I.E., GEU-FLORES, F. & HALKIER, B.A.** 2010. Biosynthesis of glucosinolates- gene discovery and beyond. *Trends in Plant Science*, **15**, 283–290.

- SONG, K., LU, P., TANG, K. & OSBORN, T.C.** 1995. Rapid genome change in synthetic polyploids of *Brassica* and its implications for polyploid evolution. *Proceedings of the National Academy of Sciences*, **92**, 7719–7723.
- SOTELO, T., SOENGAS, P., VELASCO, P., RODRIGUEZ, V.M. & CARTEA, M.E.** 2014. Identification of metabolic QTLs and candidate genes for glucosinolate synthesis in *Brassica oleracea* leaves, seeds and flower buds. *PLoS ONE*, **9**, e91428.
- Španiel, S., MARHOLD, K., PASSALACQUA, N.G. & ZOZOMOVÁ-LIHOVÁ, J.** 2011. Intricate variation patterns in the diploid-polyploid complex of *Alyssum montanum*-*A. repens* (Brassicaceae) in the Apennine Peninsula: Evidence for long-term persistence and diversification. *American Journal of Botany*, **98**, 1887–1904.
- STAMATAKIS, A.** 2014. Stamatakis - 2014 - RAxML version 8 a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 2010–2011.
- STEARNS, A.S.C.** 1989. Trade-offs in life-history evolution. *Functional Ecology*, **3**, 259–268.
- STEVENS, P.F.** 2001. Angiosperm phylogeny website.
- STÜMPPEL, N., RAJABIZADEH, M., AVCI, A., WÜSTER, W. & JOGER, U.** 2016. Phylogeny and diversification of mountain vipers (*Montivipera*, Nilson et al. 2001) triggered by multiple Plio-Pleistocene refugia and high-mountain topography in the Near and Middle East. *Molecular Phylogenetics and Evolution*, **101**, 336–351.
- SUNKAR, R. & ZHU, J.** 2004. Novel and stress-regulated microRNAs and other small RNAs from *Arabidopsis*. *The Plant Cell*, **16**, 2001–2019.
- SWIEZEWSKI, S., LIU, F., MAGUSIN, A. & DEAN, C.** 2009. Cold-induced silencing by long antisense transcripts of an *Arabidopsis* Polycomb target. *Nature*, **462**, 799–802.
- SWOFFORD, D.L.** 2002. Phylogenetic analysis using parsimony. *Options*, **42**, 294–307.
- TAKHTAJAN, A.L.** 1986. *Floristic Region of the World*. United States of America: University of California Press.
- TANG, H., BOWERS, J.E., WANG, X., MING, R., ALAM, M. & PATERSON, A.H.** 2008. Synteny and collinearity in plant genomes. *Science*, **320**, 486–488.
- TANK, D.C., EASTMAN, J.M., PENNELL, M.W., SOLTIS, P.S., SOLTIS, D.E., HINCHLIFF, C.E., BROWN, J.W., SESSA, E.B. & HARMON, L.J.** 2015. Nested radiations and the pulse of angiosperm diversification: increased diversification rates often follow whole genome duplications. *New Phytologist*, **207**, 454–467.
- TOKURIKI, N. & TAWFIK, D.S.** 2009. Chaperonin overexpression promotes genetic variation and enzyme evolution. *Nature*, **459**, 668–673.
- TREMETSBERGER, K., KÖNIG, C., SAMUEL, R., PINSKER, W. & STUESSY, T.F.** 2002. Intraspecific genetic variation in *Biscutella laevigata* (Brassicaceae): New focus on Irene Manton's hypothesis. *Plant Systematics and Evolution*, **233**, 163–181.
- VAIDYA, G., LOHMAN, D.J. & MEIER, R.** 2011. SequenceMatrix: Concatenation software for the fast assembly of multi-gene datasets with character set and codon information. *Cladistics*, **27**, 171–180.
- VAMBERGER, M., STUCKAS, H., AYAZ, D., GRACIÁ, E., ALOUFI, A.A., ELS, J., MAZANAIEVA, L.F., KAMI, H.G. & FRITZ, U.** 2013. Conservation genetics and phylogeography of the poorly known Middle Eastern terrapin *Mauremys caspica* (Testudines: *Geoemydidae*). *Organisms Diversity and Evolution*, **13**, 77–85.
- VAN ZANDT, P.A.** 2007. Plant defense, growth, and habitat: A comparative assessment of constitutive and induced resistance. *Ecology*, **88**, 1984–1993.
- VANNESTE, K., MAERE, S. & PEER, Y. VAN DE.** 2014. Tangled up in two: a burst of genome duplications at the end of the Cretaceous and the consequences for plant evolution. *Philosophical Transactions of the Royal Society of Biological Sciences*, **369**, 20130353.

## References

- VELASCO, P., SOENGAS, P., VILAR, M., CARTEA, M.E. & RIO, M. DEL.** 2008. Comparison of glucosinolate profiles in leaf and seed tissues of different *Brassica napus* crops. *Journal of the American Society for Horticultural Science*, **133**, 551–558.
- VELCHEV.** 2015. Vol.1 Plants. In: Velchev, V., ed. *Red Data Book of the PR Bulgaria*. Sofia: Publishing House Bulgarian Academy of Sciences.
- WANDELER, P., HOECK, P.E.A. & KELLER, L.F.** 2007. Back to the future: museum specimens in population genetics. *Trends in Ecology and Evolution*, **22**, 634–642.
- WARWICK, S.I. & AL-SHEHBAZ, I.A.** 2006. Brassicaceae: Chromosome number index and database on CD-Rom. *Plant Systematics and Evolution*, **259**, 237–248.
- WARWICK, S.I., MUMMENHOFF, K., SAUDER, C.A., KOCH, M.A. & AL-SHEHBAZ, I.A.** 2010. Closing the gaps: phylogenetic relationships in the Brassicaceae based on DNA sequence data of nuclear ribosomal ITS region. *Plant Systematics and Evolution*, **285**, 209–232.
- WEI, Z., JULKOWSKA, M.M., LALOË, J.O., HARTMAN, Y., DE BOER, G.J., MICHELMORE, R.W., VAN TIENDEREN, P.H., TESTERINK, C. & SCHRANZ, M.E.** 2014. A mixed-model QTL analysis for salt tolerance in seedlings of crop-wild hybrids of lettuce. *Molecular Breeding*, **34**, 1389–1400.
- WEN, J., PARKER, B.J. & WEILLER, G.F.** 2007. In silico identification and characterization of mRNA-like noncoding transcripts in *Medicago truncatula*. *In silico biology*, **7**, 485–505.
- WERNERSSON, R.** 2006. Virtual Ribosome—a comprehensive DNA translation tool with support for integration of sequence feature annotation. *Nucleic Acids Research*, **34**, W385–8.
- WICKHAM, H.** 2009. ggplot2: Elegant Graphics for data analysis. In New York, NY: Springer Verlag.
- WIERZBICKI, A.T.** 2012. The role of long non-coding RNA in transcriptional gene silencing. *Current Opinion in Plant Biology*, **15**, 517–522.
- WILLIAMSON, R., JOSEPHS, E.B. & PLATTS, A.E.** 2014. Evidence for widespread positive and negative selection in coding and conserved noncoding regions of *Capsella grandiflora*. *PLOS Genetics*, **10**, 1–12.
- WINDSOR, A.J., REICHEL, M., FIGUTH, A., SVATOŠ, A., KROYMANN, J., KLIEBENSTEIN, D.J., GERSHENZON, J. & MITCHELL-OLDS, T.** 2005. Geographic and evolutionary diversification of glucosinolates among near relatives of *Arabidopsis thaliana* (Brassicaceae). *Phytochemistry*, **66**, 1321–1333.
- WOLFE, K.H.** 2001. Yesterday's polyploids and the mystery of diploidization. *Nature Reviews Genetics*, **2**, 333–341.
- WRIGHT, K.M., ARNOLD, B., XUE, K., URINOVA, M., O'CONNELL, J. & BOMBLIES, K.** 2014. Selection on meiosis genes in diploid and tetraploid *Arabidopsis arenosa*. *Molecular Biology and Evolution*, **32**, 944–955.
- XIN, M., WANG, Y., YAO, Y., SONG, N., HU, Z., QIN, D., XIE, C., PENG, H., NI, Z. & SUN, Q.** 2011. Identification and characterization of wheat long non-protein coding RNAs responsive to powdery mildew infection and heat stress by using microarray analysis and SBS sequencing. *BMC Plant Biology*, **11**, 61.
- YOSHIDA, K., BURBANO, H.A., KRAUSE, J., THINES, M., WEIGEL, D. & KAMOUN, S.** 2014. Mining herbaria for plant pathogen genomes: Back to the future. *PLoS Pathogens*, **10**, 6–11.
- YOSHIDA, K., SCHUENEMANN, V.J., CANO, L.M., PAIS, M., MISHRA, B., SHARMA, R., LANZ, C., et al.** 2013. The rise and fall of the *Phytophthora infestans* lineage that triggered the Irish potato famine. *eLife*, **2013**, 1–25.
- YU, Y., HARRIS, A.J., BLAIR, C. & HE, X.** 2015. RASP (Reconstruct Ancestral State in Phylogenies): A tool for historical biogeography. *Molecular Phylogenetics and Evolution*, **87**, 46–49.

- ZEDANE, L., HONG-WA, C., MURIENNE, J., JEZIORSKI, C., BALDWIN, B.G. & BESNARD, G.** 2016. Museomics illuminate the history of an extinct, paleoendemic plant lineage (*Hesperelaea*, Oleaceae) known from an 1875 collection from Guadalupe Island, Mexico. *Biological Journal of the Linnean Society*, **117**, 44–57.
- ZHANG, J.** 2003. Evolution by gene duplication: an update. *Trends in Ecology and Evolution*, **18**, 292–298.
- ZHANG, J., MUJAHID, H., HOU, Y., NALLAMILLI, B.R. & PENG, Z.** 2013. Plant long ncRNAs: A new frontier for gene regulatory control. *American Journal of Plant Sciences*, **4**, 1038–1045.
- ZHANG, Y.-C. & CHEN, Y.-Q.** 2013. Long noncoding RNAs: New regulators in plant development. *Biochemical and Biophysical Research Communications*, **436**, 111–114.
- ZHANG, Y., HUAI, D., YANG, Q., CHENG, Y., MA, M., KLIEBENSTEIN, D.J. & ZHOU, Y.** 2015. Overexpression of three glucosinolate biosynthesis genes in *Brassica napus* identifies enhanced resistance to *Sclerotinia sclerotiorum* and *Botrytis cinerea*. *PLoS ONE*, **10**, 1–17.
- ZUKER, M. & STIEGLER, P.** 1981. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Research*, **9**, 133–148.



## Summary

The plant family Brassicaceae (or crucifers) is an economically important group that includes many food crops (e.g. cabbages and radishes), horticultural species (e.g. *Draba*, *Iberis*, *Lunaria*), and model plant species (particularly *Arabidopsis thaliana*). Because of the fundamental importance of *A. thaliana* to plant biology, it makes the Brassicaceae an ideal system for comparative genomics and to test wider evolutionary, ecological and speciation hypotheses. One such hypothesis is the 'Whole Genome Duplication Radiation Lag Time' (WGD-RLT) model for the role of polyploidy on the evolution of important plant families such as the Brassicaceae. The WGD-RLT model indicates a higher rate of diversification of a core-group compared to its sister group, due to a lag time after a whole genome duplication event that made it possible for novel traits and/or (geo-) ecological events to increase the core groups diversification rate.

*Aethionema* is the species-poor sister genus of the core Brassicaceae and hence is at an important comparative position to analyse trait and genomic evolution of the species-rich core group. *Aethionema* species occur mainly in the western Irano-Turanian region, which is concordantly the biodiversity hotspot of the Brassicaceae family. Moreover comparing *Aethionema* to the Brassicaceae core group can help us to understand and test the 'WGD-RLT' model. However, to be able to do so we first need to know more about *Aethionema*. In this thesis, I investigated various levels of evolutionary change (from macro, to micro to trait evolution) within the genus *Aethionema*, with a major focus the emerging model species *Aethionema arabicum*.

Next generation sequencing has made it possible to use the genomes of many species in a comparative framework. However, the formation of proteins and enzymes, and in the end the phenotype of the whole plant, relies on transcription from particular regions of the genome including genes. Hence, the transcriptome makes it possible to assess the functional parts of the genome. Gene regulatory elements like promoters and long non-coding RNAs function as regulators of gene expression and thus are involved in increasing or decreasing transcription. In Chapter 2 I used the transcriptome of four different *Aethionema* species to understand the lineage specificity of these long non-coding RNAs. Moreover in a comparison with the Brassicaceae core group and the sister family of the Brassicaceae, Cleomaceae, I show that although the position of long non-coding RNAs can be conserved, their sequences do not have to be.

Most of the *Aethionema* species occur in the Irano-Turanian region, a politically unstable region, making it hard for scientist to collect from. However the natural history collections made throughout the last centuries are a great resource. Combing these collections with the newest sequencing techniques, e.g. next generation sequencing, have allowed me to infer the phylogeny of ~75% of the known *Aethionema* species in a time calibrated and historical biogeographical framework. I was able to establish that *Aethionema* species likely originated from the Anatolian Diagonal and that major geological events like the uplift of the Turkish and Iranian plateau have had a hand in their speciation (Chapter 3).

To examine species-level processes I sequenced and analysed transcriptomes of eight *Ae. arabicum* accessions coming from Cyprus, Iran and Turkey to investigate population structure,

## Summary

genetic diversity and local adaptation (Chapter 4). The most prominent finding was a ploidy difference between the Iranian and Turkish/Cypriot lines, whereby the former were tetraploid and the latter diploid. The tetraploid Iranian lines seem to have one set of alleles from the Turkish/Cypriot gene-pool, though we do not know where the other alleles come from. In addition to the differences in ploidy level there are also differences in glucosinolate defence compounds between these two populations (Iranian vs. Turkish/Cypriot), with the Iranian lines lacking the diversity and concentration of indolic glucosinolates that the Turkish/Cypriot lines have. This chapter serves as a good resource and starting point for future research in the region, maybe by using the natural history collections that are at hand.

Glucosinolates (i.e. mustard oils) are mainly made by Brassicales species, with their highest structural diversity in the Brassicaceae. In Chapter 5, I examined two *Ae. arabicum* lines (CYP and TUR) and their recombinant inbred lines to assess glucosinolate composition in different tissues and throughout the plants development. The levels of glucosinolates in the leaves changed when *Ae. arabicum* went from vegetative to a reproductive state. Moreover, a major difference in glucosinolate content (up to 10-fold) between CYP and TUR indicates a likely regulatory pathway outside the main glucosinolate biosynthesis pathway. Multi-trait and multi-environment QTL analyses based on leaves, reproductive tissues and seeds identified a single major QTL. Fine mapping this region reduced the interval to only fifteen protein coding genes, including the two most intriguing candidates: *FLOWERING LOCUS C (FLC)* and the sulphate transporter *SULTR2;1*. These findings show an interesting correlation between development and defence.

Finally, Chapter 6 gives a final discussion of this thesis and its results. It brings the different topics together, puts them in a bigger picture and looks forward into new research possibilities.



## Acknowledgements

Here it is. My PhD-thesis. But it would not be here if it was for the mental, emotional and practical support of many people. Thanks to you all! There are a number of people who I would like to mention in particular, in a fairly random order.

Hey boss, Eric. Thank you for your trust in me and belief in my qualities, some of which I did not know I had. The many seemingly 'nonsense' moments about YouTube videos are worth their time in gold. I enjoyed the learning curve and the freedom that comes with having you as a boss and a friend. I do believe that the freedom encourages creativity, at least for me. It was also great to get to know your lovely Paige, Esme and Graham. I hope that we will stay in touch!

I had the pleasure to work with people from all over the world. Pat Edger and Chris Pires, I'll never forget my brief time at your lab in Missou, it was great! I have also really appreciated your quick responses to my various questions throughout the years. Ihsan Al-Shehbaz: you are the grandfather of this project. I hope that it has met your expectations. Please give your beautiful Mona a big hug from me. Klaus Mummenhoff: keep up all those fieldwork trips and I hope that we will continue our collaborations in the future. Hamid Moazzeni and Michiel Reichelt: Thank you for your help with getting samples or analysing them. Stephen Wright: It took some long and hard work, but in the end we made it, the popgen chapter is there! Thank you for your help and the skype meetings.

Everyone from Biosystematics: thank you for your warm hearts. Rens en Robin, jullie blijven toch soort van erbij horen. Bedankt voor alle praatjes, koffies en biertjes door de jaren heen. Dag Frank, het was gezellig! Roel: Ik hoop echt dat alles wat ik van je geleerd heb in de Pyreneeën ergens blijft hangen. Freek: Wat heerlijk om te kunnen brainstormen en na te denken over analyses (en het leven!). Heel erg veel geluk met Onkamon, ze fleurt je op ☺. Andre: Je bent er bijna! Je kunt het! Je hebt al zoveel gepresteerd in je leven! Nina: Spring-in-'t-veld, heerlijk dat je erbij bent gekomen! Carla: Het was kort maar krachtig, ik hoop je in de toekomst vaak te zien, want we hebben zeker lol samen. Laurie and Lydia: time went by to quickly, but I guess that, that happens when you have fun. Phuong: You are one of the most accurate people I know and you thought me to be more critical about my results. You have had an amazing year and I hope that we can collaborate and enjoy life in the future.

Lieve Wilma, het is vaak gezegd maar het is ook echt waar: alles loopt behoorlijk vast als jij er niet bent. Je was er altijd als ik mijn hart wilde luchten en gaf me goede raad. Dank je wel. Mi alma saluda tu alma.

Lars, hallo!!!! Ah nee hè... Ik wil geen bord! De reizen van en naar de Pyreneeën, al dan niet met airco. De wisselende jaren daar, de slechte en goede aioli's en de zakjes die blijven liggen tijdens het omhoog wandelen in de bergen. Logeren bij jullie en het leren kennen van je mooie Griete en de mannen Job, Korneel en Lucas! Alle gesprekken bij de koffie(automaat) over alles en meer. Zullen we daar gewoon mee doorgaan? Net zo makkelijk toch? Tijd is namelijk een emotie en die kan je uitzetten.

## Acknowledgements

Sara Jacobientje... hèhè... noemt iemand jou ook echt zo? Zal je vast niet bevallen :P. Je bent mooi, lief, intelligent en kunt alle borst-kloppende mannen aan. Laat je niets anders wijs maken! De avonturen zijn teveel om op te noemen, studenten, data-analyses, samen papers schrijven, winkelen, breien, etc. Laten we ervoor zorgen dat ze doorgaan.

Mijn studenten, Sara, Marco, Kim, Emmy, Feia en Frank, bedankt voor jullie geduld en het accepteren dat de wetenschappelijke waarheid er niet altijd uitziet zoals je wilt of verwacht. Marco: Ik ben heel erg blij met onze hervonden vriendschap! De biertjes moeten blijven komen, anders blijven we nooit nuchter.

Dan zijn er nog mensen die niet bij Biosystematiek zaten maar wiens steun onontbeerlijk was voor het proces.

Gerard, Ludek, Harold en Thomas: de basis is op de UvA gelegd. Veldwerk en het leren verpotten vermengd met een goede dosis humor en koffie. Bedankt vooral voordat laatste, hebben jullie mij toch echt leren drinken en ik heb er nog steeds profijt van. Bakkie doen?

Atje, Joop, Ariane, Sara en Kamiel: Bedankt voor de etentjes. Ik vind het heel fijn dat jullie mij hebben opgenomen in de familie. Joop: Bedankt voor het doorspitten van mijn proefschrift, ik vind het heel leuk en waardevol dat jij dat hebt willen doen!

Pippa en Martha: Delen van dit boekje zijn tot stand gekomen in jullie huizen, met de hulp van de kat, honden en zwembaden. Heel erg bedankt voor de altijd aanwezige gastvrijheid, lieve mevrouwjes!

The EPS girls: Hanna, Lot, Lotte and Magda. Let's stick together, wherever we are!

Rosa, Juul, Lotte, Hanna en Esther: De maandagen zijn voorbij, en de meiden blijven. We keep being fierce but beautiful! De diners, showers, koffietjes en lunches, ze zijn onmisbaar voor mij.

Hanna, Hedwig en Bram: Sinds het eerste studiejaar een vaste groep en ook al wonen we ondertussen iets verder van elkaar, de band is er niet minder op geworden. De gezamenlijke etentjes zijn heerlijk, zeker nu we bijna verdriedubbeld zijn in aantal!

Fijne zachte Isis, mag ik weer met Sint achter de kassa? Want wat moet een mens zonder? En zal je me wel altijd de goede vragen op de goede momenten stellen? Want ook daar kan ik niet zonder. Van jou heb ik geleerd om op mensen af te stappen, hoe eng dat ook is en hoe vaak dat ook mislukt. Je hebt me laten zien hoe een leven kan veranderen en variëren, en hoe elke ervaring een andere kant heeft. En het is elke keer een feest om je mooie dochters Bente en Josephine mee te maken.

Ana, Guido, Eylard, Susan, Rogier, Lisette, Jannes, Josefien, Ruben en Gusta: Bedankt dat jullie me hebben opgenomen in de Barleaus-familie. Ik vind elke keer dat we elkaar zien een warm bad waarin we elkaar accepteren inclusief onze gekkigheden en dankzij de verschillen. Gusta: Onze ontdekking van lunch-plekken in Amsterdam moet toch echt wel beter... Denk je niet?

De Club van Krikkestijn, Hedwig, Ella en Renske. Mag het een onsje meer zijn? Bedankt voor er zijn in de slechte en goede tijden. Jullie zijn voor mij een baken op moeilijke en vreugdevolle momenten en levensdingensen... Dankzij jullie houd ik mijn voeten aan de grond. Ik beschouw jullie als een soort van zusjes, maar dan wel de leukere versies :P. Ik kan niet zonder jullie.

I would not have been here if it was not for my parents, physically of course, as well as mentally. You are, for me, the best examples of carrying on with a major smile on your face, although life might not give you what you thought or dreamt about. If I have one thousands of your strength I will be fine this life-time. Moreover you raised me with the beauty of nature, which seems to have become the basis for my interest in biodiversity and evolution.

Attila, hey schatje! Ik weet niet hoe ik je moet bedanken. Voor de knuffels en de opkrikkende woorden als de put eindeloos was. En je geduld als de euforie mij in een woordenwaterval veranderde. Je bent er als mijn codes niet lukken, er zijn heerlijke pizza's, nassies, wraps en andere voedsels na een dag lang werken. En als kers op de tiramisu heb jij ook een behoorlijke bijdrage geleverd aan hoe dit boekje eruit ziet. Je leert mij ambitieus te zijn, doelen op te stellen en er achteraan te gaan. Ik kan mijn leven niet zonder jou voorstellen, met je gevoel voor humor en je scherpe blik op de wereld. Oh, en ik denk dat het gelukt is. Lieve kus.

Laters allemaal.

PS: NS bedankt voor het forenzen. In de afgelopen vier jaar heb ik nauwelijks echt vervelende vertragingen gehad.

## List of publications

T.-P. Nguyen, E. van der Berg, **S. Mohammadin**, A. Platts, C.S. Christodoulou, M.E. Schranz. *Filling the gap in the Brassicaceae comparative genomics: the genetic map and genomic blocks of Aethionema arabicum*. *In prep.*

**S. Mohammadin**, K. Peterse, S.J. van de Kerke, L.W. Chatrou, A.A. Dönmez, K. Mummenhoff, J. C. Pires, P. P. Edger, I.A. Al-Shehbaz, M.E. Schranz. *Anatolian origins and diversification of Aethionema, the sister lineage of the core Brassicaceae*. *American Journal of Botany*, *under revision*.

**S. Mohammadin**, W. Wang, T. Liu, H. Moazzeni, K. Ertugrul, T. Uysal, J.C. Pires, P. P. Edger, S.I. Wright, M.E. Schranz. *Genome-wide nucleotide diversity and associations with geography, ploidy level and glucosinolate profiles in Aethionema arabicum (Brassicaceae)*. *PLOS ONE under review*.

**S. Mohammadin**, T.-P. Nguyen, M.S. van Weij, M. Reichelt, J. Gershenzon, M.E. Schranz. *Major multi-trait quantitative trait locus controls glucosinolate content across developmental stages of Aethionema arabicum (Brassicaceae)*. *Frontiers in Plant Sciences, under review*.

S. J. van de Kerke, B. Shrestha, T.A. Ruhlman, M.-L. Weng, R.K. Jansen, C.S. Jones, C.D. Schlichting, **S. Mohammadin**, M.E. Schranz, F.T. Bakker. *Plastome based phylogenetics in Pelargonium*. *Molecular Phylogeny and Evolution, under review*.

T. Lenser, K. Graeber, Ö.S. Cevik, N. Adigüzel, A.A. Dönmez, M. Ketterman, S. Mayland-Quelhorst, **S. Mohammadin**, T.-P. Nguyen, F. Rumpler, C. Schulze, K. Sperber, T. Steinbrecher, N. Wiegard, M. Strnad, O. Mittelstein Scheid, S.A. Rensing, M.E. Schranz, G. Theißen, K. Mummenhof, G. Leubner-Metzger (2016). *Aethionema arabicum as a model system for studying developmental control and plasticity of fruit and seed dimorphism*. *Plant physiology*, 2016; pp.00838.2016.

**S. Mohammadin**, P.P. Edger, J.C. Pires, M.E. Schranz (2015). *Positionally-conserved but sequence-diverged: identification of long non-coding RNAs in the Brassicaceae and Cleomaceae*. *BMC Plant Biology*; 15;1;217-229

F.T. Bakker, D. Lei, J. Yu, **S. Mohammadin**, Z. Wei, S.J. van de Kerke, B. Gravendeel, M. Nieuwenhuis, M. Staats, D.E. Alqueza-Planas, R. Holmer (2015). *Herbarium genomics: plastome sequence assembly from a range of herbarium specimens using an Iterative Organelle Genome Assembly pipeline*. *Biological Journal of the Linnean Society*. 2010;1-11

M.E. Schranz, **S. Mohammadin**, P.P. Edger (2012). *Ancient whole genome duplications, novelty and diversification: the WGD radiation lag-time model*. *Current opinion in plant biology*. 15;2;147-153.

## About the author

Setareh Mohammadin was born in Teheran, Iran, on 21 March 1987. She moved with her parents to the Netherlands in 1993. In 2005 she received her secondary school diploma from the Spinoza Lyceum in Amsterdam.

Her academic studies started immediately thereafter at the University of Gent in Belgium, and were continued at the University of Amsterdam where she accordingly obtained her Bachelor of Science in Biology in 2009.

During her time in Amsterdam, Setareh was an active member of the Education Committee that is responsible for the continuous evaluation of the quality of the offered Biology education at the University. In addition, she organised two field excursions: to Aachen (Germany) and to Oland (Sweden). The last trip gave all her fellow students the opportunity to conduct fieldwork for their Bachelor thesis. Moreover to expand her academic horizon, Setareh enrolled in several extra-curriculum honours classes varying from quantum physics to philosophy.



She started the Master “Ecology and Evolution” at the University of Amsterdam in 2009, and carried out three internships during that time. The first internship entitled “*Effects of invasive plants on the invertebrate community of surrounding species*” was conducted under the supervision of Thomas van Hengstum in the lab of Dr Gerard Oostermeijer. Dr. Eric Schranz supervised the second internship, where Setareh worked on the genetic diversity of *Aethionema*. The last research project took place at the University of Oxford in the lab of Dr Roosa Leimu, and had as title: “*Germination, priming, preference, poop: The effect of herbivory and herbivore odours on germination and priming of Brassica oleracea and Brassica nigra*”. During her master, Setareh was the secretary of the Education Board Biology, teaching assistant for several Biology bachelor courses at the University of Amsterdam, and tutor in the first year ecology course at Pembroke College of the University of Oxford.

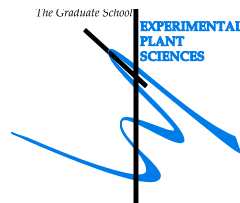
Finishing her master education in 2012, Setareh started as a PhD candidate in the Biosystematics group at Wageningen University and Research. The presented research was carried out under the supervision of Prof. Dr Eric Schranz. During her PhD, Setareh was an active member of the graduate school of Experimental Plant Sciences (EPS) and the Wageningen PhD Council (WPC).

## Education statement

### Education Statement of the Graduate School

#### Experimental Plant Sciences

Issued to: Setareh Mohammadin  
 Date: 11 April 2017  
 Group: Biosystematics  
 University: Wageningen University & Research



	<u>date</u>
<b>1) Start-up phase</b>	
▶ <b>First presentation of your project</b>	
<i>Title: Brassicaceae: crown-group vs. sister clade</i>	19 Mar 2013
▶ <b>Writing or rewriting a project proposal</b>	
▶ <b>Writing a review or book chapter</b>	
▶ <b>MSc courses</b>	
▶ <b>Laboratory use of isotopes</b>	
<i>Subtotal Start-up Phase:</i>	<i>1.5 credits*</i>
<b>2) Scientific Exposure</b>	
▶ <b>EPS PhD student days</b>	
Joint European Retreat of PhD students in Experimental Plant Sciences, Ghent, Belgium	23-26 Jul 2013
PhD student day 2013, Leiden, NL	29 Nov 2013
PhD student day 'Get2Gether', Soest, NL	29-30 Jan 2015
Joint European Retreat of PhD students in Experimental Plant Sciences, Paris, France	10-13 Jul 2015
PhD student day 'Get2Gether', Soest, NL	28-29 Jan 2016
▶ <b>EPS theme symposia</b>	
Theme 3 symposium 2013, Amsterdam, NL	21 Mar 2013
Theme 4 symposium 2013, Wageningen, NL	13 Dec 2013
Theme 4 symposium 2014, Wageningen, NL	03 Dec 2014
Theme 4 symposium 2015, Amsterdam, NL	15 Dec 2015
▶ <b>Lunteren days and other National platforms</b>	
Annual meeting 'Experimental Plant Sciences', Lunteren, NL	22-23 Apr 2013
Annual meeting 'Experimental Plant Sciences', Lunteren, NL	14-15 Apr 2014
Annual meeting 'Experimental Plant Sciences', Lunteren, NL	13-14 Apr 2015
Annual meeting 'Experimental Plant Sciences', Lunteren, NL	11-12 Apr 2016
▶ <b>Seminars (series), workshops and symposia</b>	
Prof. Dr. Rampal S. Etienne (RUG) 'A conceptual and statistical framework of adaptive dynamics'	05 Nov 2012
Prof. Dr. Thomas Mitchell Olds 'Strong selection on the genes controlling complex traits in complex environments'	10 Dec 2012
WEES Seminar: Dr. Bertus Beaumont 'Adaptive Radiation, Flagella and the Evolution of Biological Complexity'	24 Jan 2013
Prof. Dr. Marcus Koch 'Arabidopsis hybrid speciation - rare exception or simply overlooked?'	14 Feb 2013
Dr. Tim Sharbel 'The dynamics of asexual genome evolution and candidate apomixis factors in the genus Boechera (Brassicaceae)'	20 Feb 2013
Prof. Dr. Ortrun Mittelsten Scheid (Gregor Mendel Institute), 'Genetics and epigenetics: a complex relationship'	19 Nov 2014
National PhD day, The Hague, Netherlands	21 Nov 2014
Prof. Dr. George Coupland (MPI Cologne), 'Seasonal flowering in annual and perennial plants'	19 Jan 2015
Prof. Dr. Yves van de Peer (VIB Ghent), 'The evolutionary significance of gene and genome'	03 Feb 2015
Dr. Michael D. Pirie (Univ. Mainz), 'Inferring species trees given coalescence and reticulation'	18 March 2015
Prof. Dr. Hanna Kokko (Evolutionary Ecology, University of Zurich, Switzerland), 'Males exist. Does it matter?'	19 March 2015
Prof. Dr. Jeff Doyle (Cornell USA), 'Polyploidy in wild relatives of soybean and other legumes: systematics, comparative and functional genomics, and nodulation'	12 May 2015
Dr. Siobhan Brady (UC Davis, USA), 'Regulation of root morphogenesis in tomato species in the face of a changing environment'	09 Sep 2015
Prof. Dr. David Salt (Univ Aberdeen, UK), 'Landscape ionomics: The functional genomics of ecologically adaptive ionic variation'	27 Oct 2015
Dr. Rossana Henriques (CRAG, Spain), 'And yet they oscillate: functional analysis of circadian long'	16 Nov 2015
Dr. Alain Goossens (VIB, Ghent, Belgium), 'How jasmonates provide the key to harness plant'	08 Dec 2015
Koos Biesmeijer (Naturalis Biodiversity Center) 'On bees, pollination and food security'	18 Dec 2014
Prof. Dr. Jane Parker (Resistance pathway dynamics in plant immunity, MPI Cologne) 'Plant intracellular immunity: evolutionary and molecular underpinnings'	21-Jan-16
Dr. Olivier Hamant (ENS Lyon, France), 'How do plants read their own shape?'	16 Mar 2016
Dr. Sophie Nadot (CNRS Paris, France), 'Perianth evolution in Ranunculaceae: are petals ancestral in WURomics: Technology- Driven Innovation for Plant Breeding'	20 May 2016
	16 Dec 2016

## Education certificate

▶ <b>International symposia and congresses</b>		
New Model Systems for Linking Evolution and Ecology, Heidelberg, Germany		01-04 May 2013
Plant Genome Evolution, Amsterdam, The Netherlands		08-10 Sep 2013
Regulatory and Non-coding RNAs, Cold Spring Harbor, USA		26-30 Aug 2014
Population Genetics Group Meetin, Sheffield, UK		06-09 Jan 2015
Plant Genome Evolution, Amsterdam, The Netherlands		08-08 Sep 2015
New Model Systems for Linking Evolution and Ecology, Heidelberg, Germany		08-11 May 2016
▶ <b>Presentations</b>		
Poster - PhD EPS retreat, Ghent, Belgium		23-26 Jul 2013
Talk - EPS theme 4 Symposium		13 Dec 2013
Poster - Regulatory and Non-coding RNAs, Cold Spring Harbor, USA		26-30 Aug 2014
Talk - Population Genetics Group Meeting, Sheffield, UK		06-09 Jan 2015
Talk - It doesn't run in the family: Aethionemeae as a new model, ALW Lunteren 2016		11-12 Apr 2016
Talk - Aethionemeae Popeye genomics: The long non-coding RNAs, PhD EPS retreat, Paris, FR		10-13 Jul 2015
Talk - ERA-CAPS consortium meeting, Marburg, Germany		19-21 Nov 2015
Poster - New Model Systems for Linking Evolution and Ecology, Heidelberg, Germany		08-11 May 2016
▶ <b>IAB interview</b>		
▶ <b>Excursions</b>		
	<i>Subtotal Scientific Exposure:</i>	<i>24.8 credits*</i>
<b>3) In-Depth Studies</b>		<u><i>date</i></u>
▶ <b>EPS courses or other PhD courses</b>		
Postgraduate course 'Bio-informatics: a Users Approach'		04-08 Mar 2013
PhD course 'An Introduction to Phylogenetics and Population Genetics with R' (Transmitting		15-19 Apr 2013
PhD course 'Current Trends in Phylogenetics'		14-18 Oct 2013
Postgraduate Course 'Transcription Factors and Transcriptional Regulation'		17-19 Dec 2013
Postgraduate course 'Bayesian Statistics'		20-21 Oct 2014
▶ <b>Journal club</b>		
Journal club Biosystematics group		2013-2016
▶ <b>Individual research training</b>		
	<i>Subtotal In-Depth Studies:</i>	<i>9.1 credits*</i>
<b>4) Personal development</b>		<u><i>date</i></u>
▶ <b>Skill training courses</b>		
ExPectationS Day (EPS Career day): Creative Thinking		01 Feb 2013
High Impact Writing		18-21 Nov 2013
Scientific Writing		Feb-Apr 2015
▶ <b>Organisation of PhD students day, course or conference</b>		
▶ <b>Membership of Board, Committee or PhD council</b>		
	<i>Subtotal Personal Development:</i>	<i>4.8 credits*</i>
	<b>TOTAL NUMBER OF CREDIT POINTS*</b>	<b>40.2</b>

Herewith the Graduate School declares that the PhD candidate has complied with the educational requirements set by the Educational Committee of EPS which comprises of a minimum total of 30 ECTS

\* A credit represents a normative study load of 28 hours of study.

This work was supported by the grants from NWO Vernieuwings Impuls VIDI (Grant number: 864.10.001) and (849.13.004), the latter as part of the ERA-CAPS "SeedAdapt" consortium project ([www.seedadapt.eu](http://www.seedadapt.eu)). The work was carried out at the Biosystematics Group, Wageningen University, The Netherlands.

Cover design: A.M. Houtkooper

Printed by: DIGIFORCE