Plant Systematics and Evolution
https://doi.org/10.1007/s00606-018-1494-3

ORIGINAL ARTICLE

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

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

Received: 4 October 2017 / Accepted: 3 January 2018

Abstract

The genus *Aethionema* is sister to the core Brassicaceae (including *Arabidopsis thaliana*) and thus has 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 and 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. Sixteen percent of the genes in this region were chaperonins involved in protein folding. 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. We showed that different *Ae. arabicum* individuals have different ploidy levels depending on their location (Iranian versus Turkish/Cypriot). Moreover, these differences between the populations are also shown in their defence compounds.

Keywords *Aethionema* · Brassicaceae · Genome-wide diversity · Glucosinolates · Irano-Turanian · Polyploidy

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.) Andrz. 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). *Aethionema 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 (Bibalani 2012; Lenser et al. 2016). On these steep slopes with a lack of covering vegetation, *Ae. arabicum* is exposed to strong UV light, the heat of the summer and the cold of the winter. 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 (Imbert 2002; Lenser et al.

Handling editor: Marcus Koch.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00606-018-1494-3) contains supplementary material, which is available to authorized users.

✉ M. Eric Schranz
eric.schranz@wur.nl

[1] Biosystematics, Plant Science Group, Wageningen University and Research, Wageningen, The Netherlands

[2] Department of Ecology and Evolutionary Biology, University of Toronto, Toronto, ON, Canada

[3] Department of Botany, Research Center for Plant Sciences, Ferdowsi University of Mashhad, Mashhad, Iran

[4] Department of Biology, Faculty of Science, Selcuk University, Konya, Turkey

[5] Department of Forests, Ministry of Agriculture, Rural Development and Environment, Nicosia, Cyprus

[6] Department of Horticulture, Michigan State University, East Lansing, MI, USA

[7] Division of Biological Sciences, University of Missouri, Columbia, MO, USA

Published online: 09 February 2018

2016). The short life cycle and the heterocarpic phenotype are likely adaptations to the unpredictable local growth conditions of *Ae. arabicum.*

The genus *Aethionema* shares the typical cruciferous traits, e.g. the methionine-derived glucosinolates (i.e. mustard oils, GLS) with the Brassicaceae core group (Hofberger et al. 2013; Edger et al. 2015). GLS are a novel suite of metabolites developed within the Brassicales as chemical defences against pathogens and herbivores (Halkier and Gershenzon 2006; Hofberger et al. 2013; Edger et al. 2015). Glucosinolates are derived from a basic sugar molecule and an amino acid with a side chain that can be elongated with carbon molecules (Halkier and Gershenzon 2006; Redovnikovic et al. 2008). In addition to their function as defence compounds, GLS can also act as attractants and are economically used for their mustard flavour and anti-carcinogenic activity (Halkier and Gershenzon 2006). Although all Brassicales contain GLS, the At-alpha WGD event and the arms race between Brassicaceae and its Pieridae herbivores likely increased the GLS 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 whole genome duplication (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 (Zhang 2003; Ha et al. 2009; Soltis et al. 2009; Koenig and Weigel 2015). 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 and Weigel 2015).

Aethionema arabicum has mainly been studied in a comparative framework with the rest of the Brassicaceae (Beilstein et al. 2012; Hofberger et al. 2013; Mohammadin et al. 2015), to investigate the genome and transcriptome sequences (Haudry et al. 2013; Edger et al. 2015; Mohammadin et al. 2015), long noncoding RNAs and other conserved noncoding 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 population is required. Here we used a genome-wide approach to investigate the genomic diversity between different *Ae. arabicum* accessions sampled widely in the Irano-Turanian region. Although our sampling is small, the analysis of nucleotide polymorphism data derived from de novo transcriptome sequencing reveals two geographical groups that correlate with ploidy differences found by flow cytometric analyses. Furthermore, GLS analysis finds that the geographical clusters differ in their defensive GLS 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 (9 cm × 9 cm × 10 cm) 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: 36.58077N; 032.27649E and 37.01166N; 032.19826E), Turk3 from Elaziğ, Turkey (no GPS data available). Iran1 came from the Dizin mountains (Karaj, Iran, UTM coordinates: 36.06851N; 051.19645E) and Iran2 and Iran3 from the base of the Tochal mountain (Tehran, Iran, UTM coordinates: 36.06851N; 051.19645E). Cyp seeds were from Kato-Moni (Cyp, Cyprus, lat 35.057310 and lon 33.091832). Sampling localities are shown on the map in Fig. 1. RNA isolation, cDNA synthesis, library preparation and assembly followed the same procedure as for *Aethionema carneum* (Banks & Sol.) B.Fedtsch. (Mohammadin et al. 2015). Leaves, flowers, fruits, apical meristems from adult plants and whole seedlings were pooled separately for every line. Tissues were ground in liquid nitrogen after which RNA was isolated with the PureLink™ minikit (Ambion, Life Technologies Corporation, Carlsbad, CA, USA). RNA isolation was followed by a DNase treatment with the TURBO DNA-free™ kit (Ambion). RNA quality and quantity were checked on a 1% agarose gel stained with ethidium bromide in a 1× TBE buffer and on a NanoDrop 1000© (Thermo Fisher Scientific, Wilmington, DE, USA). We assessed whether the samples were DNA free with a PCR on 1/100 diluted samples with primers for a gene homologous to the *Arabidopsis thaliana* gene At5g42740 sugar isomerase (SIS). Two reverse primers were used: one covering an exon–exon boundary (5′CATTCCATACGCTCA ACAACC3′) and a reverse primer covering an intron–exon border (5′ATACGCTCAACAACCCCAAG3′). In both cases, the same forward primer was used (5′AACGGCTGA AACAATGCTG3′). The primers were designed using Primer3Plus (Untergasser et al. 2007). All samples were non-normalized. When the samples passed these requirements, they were dried with GenTegra™ (GenVault, Carlsbad, CA, USA) and sent to the sequencing core of the University of Missouri Colombia (USA). There the ds-cDNA libraries

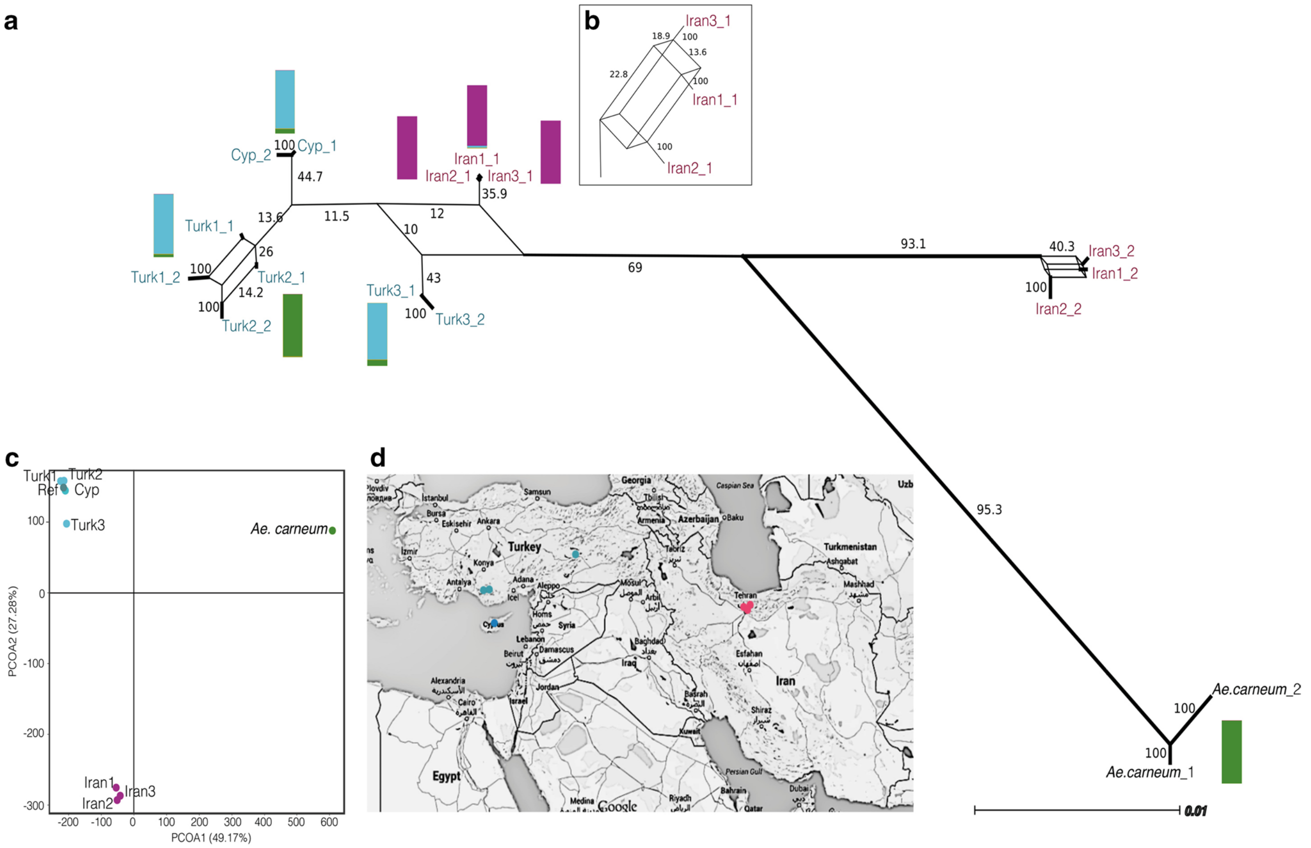

Fig. 1 Consensus network of *Aethionema arabicum* genes. **a** The network consists of 8969 bi-allelic expressed protein-coding genes of seven *Ae. arabicum* lines with *Ae. carneum* used as the out-group. The numbers along the branches show the percentage of trees supporting the branch. Naming of the tips is as follows: LocationPopulation_Allele. Hence, Turk1_1 and Turk1_2 are the two alleles of the Turk1 population, coming from Turkey. The bars along the network show the result of a STRUCTURE analysis for K = 3, only one bar per population is shown. **b** The structure of the IranX_1 allele. **c** PCOA plot showing the population clusters with the percentage of variation by the axis within brackets. **d** The sampling locations

were constructed and prepared for sequencing using the TruSeq-RNA™ kit (Illumina, San Diego, CA, USA), following manufacturers protocol. The lines were single-end sequenced with the Illumina Hiseq 2000 sequencer on 1x100 bp lanes, with 3 lines per lane resulting in 5 Gbp of reads per line. *Aethionema carneum*, 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.

Transcriptome assembly was conducted following (Mohammadin et al. 2015): reads were aligned against *Ae. arabicum* genome (Haudry et al. 2013) using NextGene V2.17 ® (SoftGenetics, State College, PA, USA) using the same parameter settings as Mohammadin et al. (2015).

Flow cytometry

In addition to the seven *Ae. arabicum* accessions 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 ClabTM, Munster, Germany) in 2011. Fully grown leaves of the same age were used (gain = 340). The nuclei were isolated by mincing the leaves with a razor blade in 1 ml of the standard isolation buffer and with 4,6′-diamidino-2-phenylindole (DAPI) (Dolezelt et al. 1998). Five-hundred microlitres chicken DNA was used as an internal standard. We also included *Brassica oleracea* as a control with a known genome size to assess and ploidy levels.

SNP calling and filtering

SNPs were called with the UnifiedGenotyper 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. GATKs UnifiedGenotyper estimates the most likely genotype and allele frequencies using a Bayesian genotype likelihood model. However, to be certain that ploidy level (see Results) did not interfere with the population structure analyses and the genome-wide diversity levels we did all downstream analyses separately for different ploidy levels.

Diploid alignments of the coding regions were made directly from the raw reads and based on the GATK output using in-house Perl scripts. Hence, 'allele_1' is the same as the *Ae. arabicum* reference and 'allele_2' containing all the mutations.

The VCF file form GATK was filtered with GQ ≥ 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.

To further validate the assignment of potential alleles for heterozygous regions, we additionally called the alleles based on read-backed phased SNPs with HapCut (Bansal and Bafna 2008). SNPs with a GQ ≤ 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 (with K = 2–5, with a 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 ('–012' 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 (Paradis et al. 2004; R Core Team 2014; Kane et al. 2013). With the exception of the network trees, 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), alignments are available upon request from the authors). A consensus network was made from all the maximum likelihood trees in SplitsTree [mean edge-weight and threshold = 0.1, (Huson and 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 and Freeling 2008). Genes from the unphased alignments with only one syntenic hit and no proxies that occurred in our gene set were used.

We used the first twenty-five genes of the phased data set 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 data set.

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 Cypriot 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 and Andolfatto 2006) for every allele type separately and averaged per allele to have a gene per individual specific descriptive. To be able to assess genome-wide patterns, we distributed the SNPs along our genetic linkage map (Nguyen et al., in prep). The ratio of non-synonymous to synonymous divergence, dN/dS, between *Ae. arabicum* and *Ae. carneum* was directly calculated from the Phylomorphorama output using the ratio of Jukes–Cantor corrected 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. dN/dS > 3 were omitted, as they might indicate not real selection but the misalignments from different members of a gene family. Sliding windows were calculated for windows of 50 cM with a step size of 8 cM 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 and Freeling 2008).

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 12-cm 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 h and ground with 4–10 2 mm glass beads. Samples that were lighter than 4 mg were pooled. Samples between 4 and 10 mg were extracted with 1 mL 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; 8.1–10 min 100% B and 10.1–13.5 min 10% B; flow 1.0 mL min^{-1}). 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.

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 out-group species *Ae. carneum* (Mohammadin et al. 2015).

The heterozygosity levels of the Iranian lines exceeded the percentage of heterozygous loci between the *Ae. arabicum* reference line and the out-group *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 Iranian lines is likely caused by (disomic) polyploidy.

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

Line	Ploidy (2×)	Heterozygosity (%)
Ae. carneum	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

The *Ae. arabicum* accessions from Turkey, Cyprus, Iran and *Ae. carneum* had an average of 38,689,493.8 (± 7951,220.4 SD) reads before quality trimming. 87.3% of the reads remained after quality trimming (33,776,390.2 ± 6958,395.1 SD). These were assembled relative to the reference line (Haudry et al. 2013) into an average of 18,881 (± 210.6 SD) contigs with an average length of 803 bp (± 3.1 SD).

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 to 22,088,876 bi-allelic sites, representing 9070 coding genes. We found 75,726 of polymorphic sites in these 9070 coding genes within *Ae. arabicum* and 860,986 SNPs in *Ae. carneum.* There were 244,184 sites heterozygous for at least one of the Iranian lines, from which 218,184 sites were heterozygous for all the Iranian lines. This accounts to: 89.35% of all the heterozygous sites and 0.98% of all bi-allelic sites. For the Turkish/Cypriot lines, 86,287 sites had at least one individual with a heterozygous locus and 38,546 sites were heterozygous for all the Turkish and Cypriot individuals. This accounts for 44.67% of all heterozygous sites and 0.17% of all bi-allelic sites. Taken together, the much higher level of heterozygosity in the polyploidy Iranian lines is most consistent with allopolyploid origins of this population.

The PCOA and STRUCTURE analyses show the clustering of the Iranian and the Turkish lines (Online Resource 1, 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 of *Ae. arabicum* is nested with other Turkish lines (Fig. 1).

Out of the 9070 coding genes with SNPs, trees were made out of 8969 unphased alignments. The consensus network of all 8969 trees shows 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 out-group species *Ae. carneum.* This separation suggests a likely allotetraploid 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 (Online Resources 2, 3).

Following the above-mentioned ploidy differences between the Iranian and Turkish/Cypriot 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 data set 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 *Ae. carneum* (Table 2). The synonymous polymorphism of the Iranian lines was much lower than the Turkish lines (π_{Iran} ~ 0.12% versus π_{Turk} ~ 1.08%, Table 2), although the Iranian lines come from a very restricted geographical region. The levels of polymorphism of non-synonymous

Table 2 Population genomic descriptive for Turkish, Cypriot and Iranian lines of *Aethionema arabicum*

Group	Site Class	N_{sites}	S	D	π	D_{xy}
Turkey + Cyprus (*n* = 4)	Synonymous	2,545,326.617	46737.5	130091.5	0.0108 (0.0156)	0.0636 (0.0338)
	Non-synonymous	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)
	Non-synonymous	8,774,195.667	4060	70775.5	0.0003 (0.0012)	0.0087 (0.0076)

Aethionema carneum was used as out-group to calculate the values for segregating sites

Averages are shown for the synonymous nucleotide diversity (*π*) and the pairwise divergence (*Dxy*) with standard deviation in brackets

N_{sites} total number of the site class, *S* number of polymorphic sites, *D* number of segregating sites

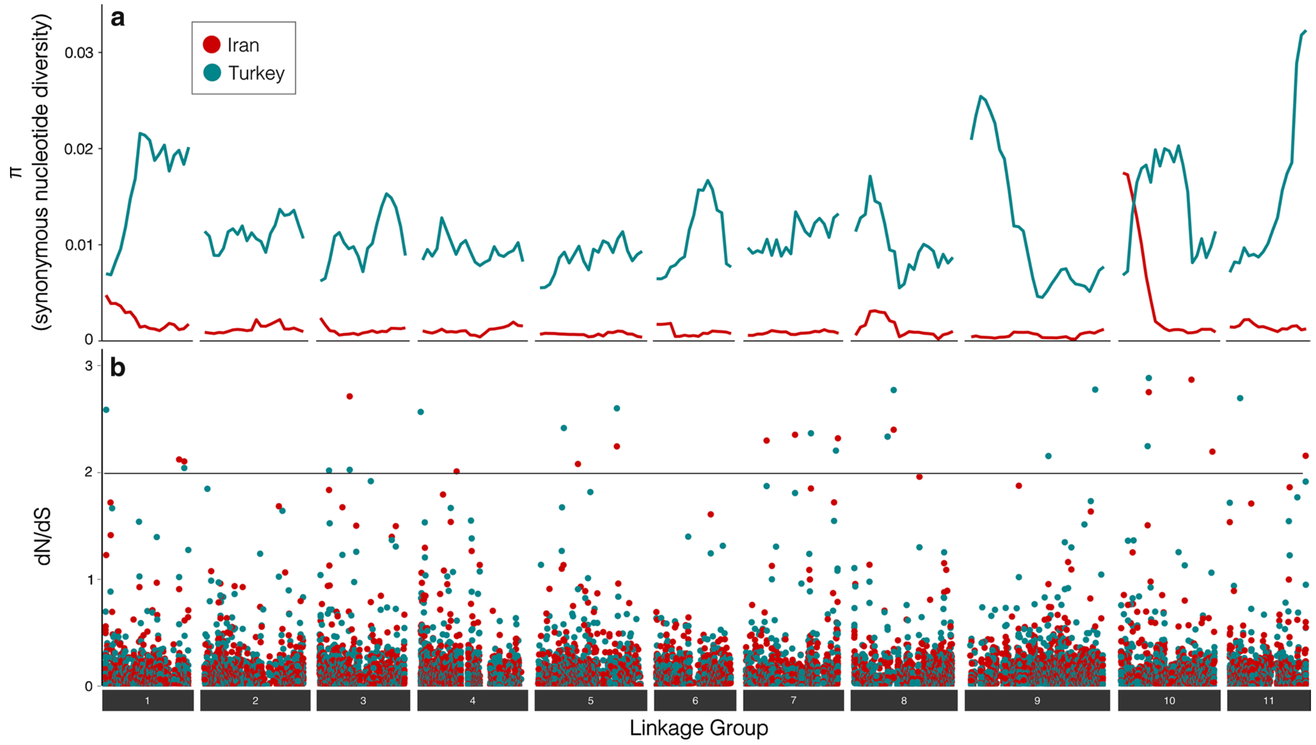

Fig. 2 Genome-wide diversity descriptive of *Aethionema arabicum*. Synonymous nucleotide diversity and the ratio of non-synonymous to synonymous divergence (dN/dS) along the linkage map (Nguyen et al. in prep.) of Iranian (red) and Turkish/Cypriot (blue) *Ae. arabicum* populations. **a** Iranian (red) and Turkish (blue) nucleotide diversity of protein-coding genes over a sliding window over 50 cM with a step size of 8 cM. **b** dN/dS ratios of Iranian (red) and Turkish (blue) protein-coding genes. Every dot represents a gene

sites were decreased by an order of magnitude for both groups (π_{Iran} ~ 0.03% versus π_{Turk} ~ 0.14%, Table 2).

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 (cut-off dN/dS < 3) were calculated between the out-group species *Ae. carneum* and the two populations (Iran and Turkey). Of the 9070 genes that passed the SNP calling restrictions, only 6762 could be mapped relative to the genome. Of the 2203 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 1616 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 10.6 cM (Fig. 2a). The synonymous polymorphism level is significantly higher here than in the rest of the transcripts (average $\pi_{syn} = 0.016 \pm 0.0096$ SD, Wilcoxon rank sum test with continuity correction, $W = 20482$, $P < 2.2e^{-16}$). Within the same 10.55-cM interval, the Turkish lines have lower but increasing synonymous substitution rate, although this does not differ from the rest of the transcripts (average $\pi_{syn} = 0.0074 \pm 0.0093$ SD, Wilcoxon Rank Sum test with continuity correction, $W = 216{,}580$, $P = 0.2937$). This region harbours 59 genes with dN/dS averages of 0.1434 (± 0.156 SD, $n = 57$) for the Iranian lines and 0.1434 (± 0.1666 SD, $n = 55$) for the Turkish lines (Online Resource 4). dN/dS values were not significantly different from the rest of the genome (Wilcoxon rank sum tests with continuity correction, Iran: $W = 238{,}220$, $P = 0.896$; Turkey $W = 240{,}400$, $P = 0.577$). However, as the dN/dS values here are below one there is a signal of purifying selection. Most of the genes which belong to an enzymatic group are involved in DNA and/or RNA binding, zinc ion binding or transcription factors (Online Resource 4). There are also genes involved in seed maturation, embryogenesis, oxygen sensing, mitochondrion inner membrane proteins and chloroplast chaperonins (Online Resource 4).

The averages of non-synonymous versus synonymous substitutions, dN/dS, were significantly higher for Turkey 0.143 (± 0.211 SD, $n = 6673$) than for Iran: 0.142 (± 0.193 SD, $n = 6762$, Wilcoxon Rank Sum test with continuity correction, $W = 23{,}378{,}000$, $P = 2.784e^{-04}$). Except for linkage group six, all the linkage groups had genes under selection with a dN/dS above 2 (Fig. 2). There were a total 25 genes with evidence for being under strong positive selection, with dN/dS values above 2 (Table 3). Four out of these 25 genes are protein-coding genes within the organelles, mitochondrion and/or chloroplasts. Six out of 24 genes are selected in both Iranian and Turkish lines. These genes encode fundamental cellular functions, e.g. RNA-binding proteins, a homolog of the mammalian regulator of apoptosis, mitochondrial importer subunits (Table 3).

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 GLS profiles varied: the Turkish and Cypriot lines have a higher diversity of GLS components compared with the Iranian lines. Out of the 89 genes of the GLS pathway that are syntenic with *Ae. arabicum* (Hofberger et al. 2013), 39 were expressed in our transcriptomes (Online Resource 5). Our filtered unphased SNP list contained 17 *Ae. arabicum* syntelogs of the GLS pathway (Online Resource 5), including genes from the entire biosynthetic pathway, ranging from the sulphatases SOT18, that are involved in the final step of GLS compound formation to the side-chain changing GOX5 (Online Resource 5) (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 π for the Turkish lines.

Discussion

Here we used a genome-wide approach and transcriptome analysis to understand the genetic diversity and selection pattern of seven *Ae. arabicum* individuals from a wide geographical range. Despite our small sample size, we show that *Ae. arabicum* lines 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 allotetraploidy or introgression. Genome-wide diversity analyses showed that from the genes that are under selection, 16% have organeller functions largely within mitochondria and/or chloroplasts. As selection happens at the level of traits, we analysed glucosinolate (GLS) defence compounds between the populations, showing a low diversity of indolic GLS in the Iranian lines while the Turkish/Cypriot lines had a higher indolic concentration and diversity.

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

Table 3 *Aethionema arabicum* genes under selection

Gene	dN/dS Iran	dN/dS Turkey	LG	Ath homolog	Function
AA46G00131	2.12	NA	1	AT2G22420	Peroxidase superfamily protein
AA46G00063	2.1	2.04	1	AT2G21390	Coatomer, alpha subunit
AA37G00130	NA	2.59	1	AT1G16250	Galactose oxidase/kelch repeat superfamily protein
AA26G00586	2.71	2.03	3	AT3G19130	RBP47B; RNA-binding protein 47B
AA26G00211	NA	2.02	3	AT3G14990	ATDJA1A; a homolog of animal DJ-1 superfamily protein
AA32G00948	2.01	NA	4	AT2G27180	Unknown protein
AA32G00024	NA	2.57	4	AT4G00752	UBX domain-containing protein
AA53G00888	NA	2.42	5	AT3G26782	Tetratricopeptide repeat (TPR)-like superfamily protein
AA53G00570	2.08	NA	5	AT1G61740	Sulphite exporter TauE/SafE family protein
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
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
AA41G00046	2.34	2.34	8	AT5G51150	Mitochondrial import inner membrane translocase subunit
AA4G00198	NA	2.15	9	AT5G64650	Ribosomal protein L17 family protein
AA21G00354	NA	2.78	9	AT2G46200	Unknown protein.
AA61G00625	2.2	NA	10	AT3G62980	Tetratricopeptide repeat (TPR)-like superfamily protein
AA61G00276	2.86	NA	10	AT3G58570	P-loop containing nucleoside triphosphate hydrolases superfamily protein
AA2G00087	NA	2.25	10	AT1G49540	ELP2; elongator protein 2
AA2G00072	2.75	2.88	10	AT1G49340	PI4 K; phosphatidylinositol 4-kinase expressed in inflorescences and shoots
AA20G00085	2.16	NA	11	AT2G20920	Protein of unknown function
AA57G00218	NA	2.69	11	AT4G19600	Encodes a cyclin T partner CYCT1;4. plays important roles in infection with cauliflower mosaic virus (CaMV)

Forty-two *Ae. arabicum* (*Gene*) genes with the values of the ratio of non-synonymous to synonymous divergence (*dN/dS Iran, dN/dS Turkey*) above 2 and the function of their *Arabidopsis thaliana* homologs (*ATH homolog*). Ordering is according to the linkage map of Nguyen et al. (in prep.)

Table 4 Glucosinolate content *Aethionema arabicum* accessions

		Iranian			Turkish			Cypriot		
		Leaves	Flowers	Fruits	Leaves	Flowers	Fruits	Leaves	Flowers	Fruits
Aliphatic	3 MTP	1.7 (1.1)	4.8	1.1 (0.7)	0.5 (0.4)	5.6	3.1 (4.1)	0.9	5.8	12.0
	3 MSOP	2.7 (1.3)	0	1.6 (1.0)	0.5 (0.4)	3.2	2.8 (3.3)	0.7	0	12.9
	3 MSOOP	0.6 (0.2)	0.3	2.0 (1.5)	0.3 (0.1)	0.4	1.1 (0.8)	1.1	0.4	2.0
	7MSOH	0 (0)	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)	0.2 (0.1)	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
	I3 M	0.1 (0)	0.1	0.1 (0)	0.1 (0)	0.1	0.3 (0.5)	0.08	0.1	0.2
Indolic	4MOI3M	0 (0)	0.1	0.1 (0)	0.2 (0.3)	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 (0.1)	0.004	0	0.4
	4OHI3M	0 (0)	0	0	0 (0)	0	0.1 (0.2)	0.002	0	0
	Total	0.1	0.2	0.2	0.4	0.3	0.8	0.116	0.2	1.0

Accessions are from Iran, Turkey and Cyprus lines of *Ae. arabicum*. Given are the glucosinolate concentrations in µmol/g with standard deviation in brackets ($n = 3$)

Samples without standard deviation had to be pooled for GS analysis 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 GLS phenotypes (Nguyen et al. in prep, (Mohammadin et al. 2017)).

The high heterozygosity levels of the tetraploid Iranian lines are consistent with hypotheses of allopolyploidy or introgression. The gene-network analyses show that one Iranian allele is incorporated into the Turkish/Cypriot cluster. Similar to *Ae. carneum,* ~ 70% of the ~ 9000 trees support the split of the second Iranian allele from the Turkish/Cypriot cluster (Fig. 1 and Online Resources 2, 3), indicating allopolyploidy. However, the low heterozygosity level of 3% of the Iranian lines indicates introgression. Introgression could also lead to tetrasomic or intermediate inheritance (Schmickl and Koch 2011). However, as our sample size is small and consists of RNA-seq data, a lot of departures from the 1:1 ratio under disomic inheritance are expected. To overcome this caveat in our methods, we would need a bigger sample size consisting of genomic sequence data. Introgression as well as allopolyploidy is common within the Brassicaceae. For example, while the diploid populations of *Arabidopsis lyrata* ssp. *petraea* and *Arabidopsis arenosa* are morphologically very distinct, their introgressed tetraploids in Eastern Europe overlap (Schmickl and Koch 2011). Allopolyploidization often follows hybridization, adapting each parental genome, such as the allotetraploid *Arabidopsis kamchatica* is a hybrid between *A. lyrata* ssp. *petraea* and *Arabidopsis halleri* (Schmickl et al. 2010). 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 (Schranz 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 of the European Alpine system (Schmickl et al. 2012). 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 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 geographical 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 and 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, biennial 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 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 and Tawfik 2009). Tokuriki and Dawfik (2009) showed that *Escherichia coli* cells without GroEL/GoES chaperons have more than a twofold decrease in non-synonymous substitution. The GroEL/GoES system is the bacterial homolog of the chaperonins; we find here that are under purifying selection.

Glucosinolate (GLS) profiles vary depending on tissue and at the species level (Kliebenstein et al. 2001; Brown et al. 2003). The GLS 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 GLS and the latter contains four different indolic compounds in much higher concentrations. Plants lacking indolic glucosinolates are more susceptible to necrotrophic fungi [67 and the references therein]. Due to our small sample size and the pooled tissues for our transcript isolation, we were not able to assess differential gene expression. The diversity of indolic compounds found in our study is mainly due to GLS side chain modifications (Sønderby et al. 2010). Although we did not find any of the GLS genes being under selection pressure, the difference found here between the GLS compounds and populations might indicate a start of local adaptation or an effect of the differences in ploidy level being expressed.

Conclusion

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 data sets generated. This resulted in the finding that the genome-wide genetic diversity was increased within the tetraploid Iranian lines containing genes encoding for chaperonins. Chaperonins are expressed when plant cells are stressed, encouraging the hypothesis that both populations deal with stressful 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.*

Acknowledgements This work was partially funded by a grant from the Netherlands Science Organization (NWO) to MES (849.13.004) as part of the ERA-CAPS "SeedAdapt" consortium project (http://www.seedadapt.eu).

Information on Electronic Supplementary Material

Online Resource 1. STRUCTURE plots showing the population clusters of *Aethionema arabicum*.

Online Resource 2. Consensus network *Aethionema arabicum.* These are 158 bi-allelic expressed protein-coding single copy genes of seven *Ae. arabicum* lines with *Ae. carneum* used as the out-group.

Online Resource 3. Consensus network of *Aethionema arabicum* 25 bi-allelic expressed genes.

Online Resource 4. Population genomic descriptive and function of *Arabidodpsis thaliana* homologs of *Aethionema arabicum* genes of the 59 'Iranian peak' genes.

Online Resource 5. Population genomic descriptive and function of *Arabidodpsis thaliana* homologs of *Aethionema arabicum* glucosinolate genes.

References

Bachtrog D, Andolfatto P (2006) Selection, recombination and demographic history in *Drosophila miranda*. Genetics 174:2045–2059. https://doi.org/10.1534/genetics.106.062760

Bansal V, Bafna V (2008) HapCUT: an efficient and accurate algorithm for the haplotype assembly problem. Bioinformatics 24:153–159. https://doi.org/10.1093/bioinformatics/btn298

Beilstein MA, Brinegar AE, Shippen DE (2012) Evolution of the Arabidopsis telomerase RNA. Frontiers Genet 3:1–8. https://doi.org/10.3389/fgene.2012.00188

Bibalani GH (2012) Investigation on flowering phenology of Brassicaceae in the Shanjan region Shabestar district, NW Iran (usage for honeybees). Ann Biol Res 6:1958–1968

Brochmann C (1993) Reproductive strategies of diploid and polyploid populations of arctic *Draba* (Brassicaceae). Pl Syst Evol 185:55–83. https://doi.org/10.1006/bojl.2001.0450

Brown PD, Tokuhisa JG, Reichelt 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*. J Chem Ecol 32:2333–2349

R Core Team (2014) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. http://www.r-project.org/

Danecek P, Auton A, Abecasis G et al (2011) The variant call format and VCFtools. Bioinformatics 27:2156–2158. https://doi.org/10.1093/bioinformatics/btr330

DePristo MA, Banks E, Poplin R et al (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43:491–498. https://doi.org/10.1038/ng.806

Dolatyari A, Vallès J, Naghavi MR, Shahzadeh Fazeli SA (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. Pl Syst Evol 299:1503–1518. https://doi.org/10.1007/s00606-013-0813-y

Dolezelt J, Greilhuber J, Lucrettiii S et al (1998) Plant Genome Size Estimation by Flow Cytometry: inter-laboratory Comparison. Ann Bot (Oxford) 82:17–26

Duarte JM, Wall PK, Edger PP et al (2010) Identification of shared single copy nuclear genes in *Arabidopsis*, *Populus*, *Vitis* and *Oryza* and their phylogenetic utility across various taxonomic levels. BMC Evol Biol 10:1–18. https://doi.org/10.1186/1471-2148-10-61

Edger PP, Heidel-Fischer HM, Bekaert M et al (2015) The butterfly plant arms-race escalated by gene and genome duplications. Proc Natl Acad Sci USA 112:8362–8366. https://doi.org/10.1073/pnas.1503926112

Gossmann TI, Song B-H, Windsor AJ et al (2010) Genome wide analyses reveal little evidence for adaptive evolution in many plant species. Molec Biol Evol 27:1822–1832. https://doi.org/10.1093/molbev/msq079

Ha M, Kim E, Chen ZJ (2009) Duplicate genes increase expression diversity in closely related species and allopolyploids. Proc Natl Acad Sci USA 106:2295–2300

Halkier BA, Gershenzon J (2006) Biology and biochemistry of glucosinolates. Annual Rev Pl Biol 57:303–333

Haudry A, Platts AE, Vello E et al (2013) An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. Nat Genet 45:891–898. https://doi.org/10.1038/ng.2684

Hofberger JA, Lyons E, Edger PP et al (2013) Whole genome and tandem duplicate retention facilitated glucosinolate pathway diversification in the mustard family. Genome Biol Evol 5:2155–2173. https://doi.org/10.1093/gbe/evt162

Huson DH, Bryant D (2006) Application of phylogenetic networks in evolutionary studies. Molec Biol Evol 23:254–267

Imbert E (2002) Ecological consequences and ontogeny of seed heteromorphism. Perspect Pl Ecol 5:13–36. https://doi.org/10.1078/1433-8319-00021

Jordon-Thaden IE, Al-Shehbaz IA, Koch MA (2013) Species richness of the globally distributed, arctic-alpine genus *Draba* L. (Brassicaceae). Alpine Bot 123:97–106. https://doi.org/10.1007/s00035-013-0120-9

Kane MJ, Emerson JW, Weston S (2013) Scalable strategies for computing with massive data. J Stat Software 55:1–19

Keeler KH (1992) Local polyploid variation in the native prairie grass *Andropogon gerardii*. Amer J Bot 79:1229–1232

Kliebenstein DJ, Kroymann J, Brown P et al (2001) Genetic control of natural variation in *Arabidopsis* glucosinolate accumulation. Pl Physiol 126:811–825. https://doi.org/10.1104/pp.126.2.811

Koenig D, Weigel D (2015) Beyond the thale: comparative genomics and genetics of *Arabidopsis* relatives. Nat Rev Genet 16:285–298. https://doi.org/10.1038/nrg3883

Lenser T, Graeber K, Cevik ÖS et al (2016) *Aethionema arabicum* as a model system for studying developmental control and plasticity of fruit and seed dimorphism. Pl Physiol 172:1691–1707. https://doi.org/10.1104/pp.16.00838

Levy-Rimler G, Bell RE, Ben-Tal N, Azem A (2002) Type I chaperonins: not all are created equal. FEBS Lett 529:1–5. https://doi.org/10.1016/S0014-5793(02)03178-2

Lyons E, Freeling M (2008) How to usefully compare homologous plant genes and chromosomes as DNA sequences. Pl J 53:661–673. https://doi.org/10.1111/j.1365-313X.2007.03326.x

Lysak MA, Mandáková T, Schranz ME (2016) Comparative paleogenomics of crucifers: ancestral genomic blocks revisited. Curr Opin Pl Biol 30:108–115. https://doi.org/10.1016/j.pbi.2016.02.001

Marhold K, Kudoh H, Pak JH et al (2010) Cytotype diversity and genome size variation in eastern Asian polyploid *Cardamine* (Brassicaceae) species. Ann Bot (Oxford) 105:249–264. https://doi.org/10.1093/aob/mcp282

McKenna A, Hanna M, Banks E et al (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20:1297–1303. https://doi.org/10.1101/gr.107524.110

Mohammadin S, Edger PP, Pires JC, Schranz ME (2015) Positionally-conserved but sequence-diverged: identification of long non-coding RNAs in the Brassicaceae and Cleomaceae. BMC Pl Biol 15:217. https://doi.org/10.1186/s12870-015-0603-5

Mohammadin S, Nguyen T-P, van Weij MS et al (2017) Flowering Locus C (FLC) is a potential major regulator of glucosinolate content across developmental stages of *Aethionema arabicum* (Brassicaceae). Frontiers Pl Sci 8:876. https://doi.org/10.3389/fpls.2017.00876

Paradis E, Claude J, Strimmer K (2004) APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20:289–290. https://doi.org/10.1093/bioinformatics/btg412

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945–959

Redovnikovic IR, Glivetic T, Delonga K, Vorkapic-Furac J (2008) Glucosinolates and their potential role in plant. Period Biol 110:297–309

Schmickl R, Koch MA (2011) *Arabidopsis* hybrid speciation processes. Proc Natl Acad Sci USA 108:141972. https://doi.org/10.1073/pnas.1104212108

Schmickl R, Jørgensen MH, Brysting AK, Koch MA (2010) The evolutionary history of the *Arabidopsis lyrata* complex: a hybrid in the amphi-Beringian area closes a large distribution gap and builds up a genetic barrier. BMC Pl Biol 10:1–18. https://doi.org/10.1186/1471-2148-10-98

Schmickl R, Paule J, Klein J et al (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. https://doi.org/10.1371/journal.pone.0042691

Schranz ME, Dobeš C, Koch MA, Mitchell-Olds T (2005) Sexual reproduction, hybridization, apomixis, and polyploidization in the genus *Boechera* (Brassicaceae). Amer J Bot 92:1797–1810. https://doi.org/10.3732/ajb.92.11.1797

Schranz ME, Mohammadin S, Edger PP (2012) Ancient whole genome duplications, novelty and diversification: the WGD Radiation Lag-Time Model. Curr Opin Pl Biol 15:147–153. https://doi.org/10.1016/j.pbi.2012.03.011

Sharbel TF, Mitchell-olds T, Dobeš C (2005) Biogeographic distribution of polyploidy and B chromosomes in the apomictic *Boechera holboellii* complex. Cytogenet Genome Res 292:283–292. https://doi.org/10.1159/000082411

Slotte T, Foxe JP, Hazzouri KM, Wright SI (2010) Genome-wide evidence for efficient positive and purifying selection in *Capsella grandiflora*, a plant species with a large effective population size. Molec Biol Evol 27:1813–1821. https://doi.org/10.1093/molbev/msq062

Soltis DE, Albert VA, Leebens-Mack J et al (2009) Polyploidy and angiosperm diversification. Amer J Bot 96:336–348. https://doi.org/10.3732/ajb.0800079

Sønderby IE, Geu-Flores F, Halkier BA (2010) Biosynthesis of glucosinolates–gene discovery and beyond. Trends Pl Sci 15:283–290. https://doi.org/10.1016/j.tplants.2010.02.005

Song K, Lu P, Tang K, Osborn TC (1995) Rapid genome change in synthetic polyploids of *Brassica* and its implications for polyploid evolution. Proc Natl Acad Sci USA 92:7719–7723

Španiel S, Marhold K, Passalacqua NG, 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. Amer J Bot 98:1887–1904. https://doi.org/10.3732/ajb.1100147

Stamatakis A (2014) Stamatakis-2014–RAxML version 8 a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. https://doi.org/10.1093/bioinformatics/btu033

Tokuriki N, Tawfik DS (2009) Chaperonin overexpression promotes genetic variation and enzyme evolution. Nature 459:668–673. https://doi.org/10.1038/nature08009

Tremetsberger K, König C, Samuel R et al (2002) Infraspecific genetic variation in *Biscutella laevigata* (Brassicaceae): new focus on Irene Manton's hypothesis. Pl Syst Evol 233:163–181. https://doi.org/10.1007/s00606-002-0189-x

Untergasser A, Nijveen H, Rao X et al (2007) Primer3Plus, an enhanced web interface to Primer3. Nucl Acids Res 35:71–74. https://doi.org/10.1093/nar/gkm306

Wickham H (2009) ggplot2: Elegant Graphics for data analysis. Springer Verlag, New York

Williamson R, Josephs EB, Platts AE (2014) Evidence for widespread positive and negative selection in coding and conserved noncoding regions of *Capsella grandiflora*. PLoS Genet 10:1–12. https://doi.org/10.1371/journal.pgen.1004622

Wright KM, Arnold B, Xue K et al (2014) Selection on meiosis genes in diploid and tetraploid *Arabidopsis arenosa*. Molec Biol Evol 32:944–955. https://doi.org/10.1093/molbev/msu398

Zhang J (2003) Evolution by gene duplication: an update. Trends Ecol Evol 18:292–298. https://doi.org/10.1016/S0169-5347(03)00033-8