Repeated rearrangement of the GSL-AOP locus contributed to spatio-temporal variation in defense chemistry in Arabidopsis

Konrad Weber
University of Copenhagen https://orcid.org/0000-0003-4090-5042
Lucasz Krych
University of Copenhagen
Meike Burow (mbu@plen.ku.dk)
University of Copenhagen https://orcid.org/0000-0002-2350-985X

Article

Keywords: Metabolic Coordination, Regulatory Networks, Biosynthesis, Glucosinolate Defense Compounds, Phylogenetic Analysis

Posted Date: December 22nd, 2020

DOI: https://doi.org/10.21203/rs.3.rs-108131/v1

Repeated rearrangement of the *GSL-AOP* locus contributed to spatio-temporal variation in defense chemistry in Arabidopsis

Konrad Weber[1,2], Lucasz Krych[3] & Meike Burow[1,2]

[1]DynaMo Center, Department of Plant and Environmental Sciences, Faculty of Science, University of Copenhagen, Thorvaldsensvej 40, 1871 Frederiksberg C, Denmark.

[2]Copenhagen Plant Science Centre, Department of Plant and Environmental Sciences, Faculty of Science, University of Copenhagen, Thorvaldsensvej 40, 1871 Frederiksberg C, Denmark.

[3]Department of Food Science, Faculty of Science, University of Copenhagen, Thorvaldsensvej 40, 1871 Frederiksberg C, Denmark.

Abstract

Plants coordinate metabolic and developmental processes with the help of genetically variable, interconnected regulatory networks. The *GSL-AOP* locus in *Arabidopsis thaliana* encodes enzymes involved in the biosynthesis of glucosinolate defense compounds and has been attributed regulatory functions e.g. in flowering time control. To correlate genetic and phenotypic variation linked to *GSL-AOP*, we conducted a phylogenetic analysis across 1135 accessions and found that the available short-read sequencing data does not fully resolve the structural diversity in the locus. We analyzed a selection of 74 accessions for glucosinolate profiles and flowering time under different conditions and acquired long-read sequence information for glucosinolate and flowering time loci. Especially in the Caucasus region, structural variation in *GSL-AOP* was associated with conditional, tissue-specific glucosinolate profiles. Variation in *FLC* among the Caucasian accessions correlated with variation in the flowering time response to vernalization, suggesting that local adaptation has shaped defense and development in an orchestrated manner.

Arabidopsis thaliana populations evolved different, but not unique strategies to secure survival and reproduction in fluctuating environments[1]. The underlying intraspecific genetic variation has revealed interconnected gene networks that integrate a multitude of environmental factors to shape complex traits like root architecture[2], nitrogen-associated metabolism[3], or pathogen resistance[4]. Recent genome sequencing studies indicated that we have underestimated the degree of genetic variation concerning gene copy numbers and chromosome-level rearrangements[5]. Accordingly, relatively little is known about the impact of locus structure and variation in non-coding DNA on fitness-related traits in *A. thaliana*.

In order to reach the transition to flowering and ensure reproductive success, the plant needs to survive and orchestrate development and defense against herbivores and biotic attackers. *A. thaliana* biosynthesizes a diverse composition of glucosinolates (GSLs), which can be activated to a range of bioactive compounds[6]. Structural variation in methionine-derived GSLs is introduced upon initial elongation of the methionine side chain by one to six methylene groups, which is determined by the allelic status of the *ELONGATION LOCUS* (*GSL-ELONG*)[7]. Secondary modifications of methylthioalkyl (MT) GSLs can further expand structure diversity, beginning with *S*-oxygenation by flavin-containing monooxygenases to methylsulfinylalkyl (MS) GSLs[8,9]. Short chain (SC) MS GSLs are the substrates for two 2-oxoglutarate–dependent dioxygenases, namely AOP2 and AOP3, encoded by a tail-to-tail tandem array in the ALKENYL HYDROXYALKYL PRODUCING locus *GSL-AOP* (Fig. 1a). The two enzymes convert the same MS precursors to different products, i.e. to alkenyl GSLs (AOP2) and hydroxyalkyl GSLs (AOP3) (Extended Data Fig. 1). AOP2 and AOP3 products can be further converted to OH-alkenyl or benzoyloxyalkyl GSLs, respectively[9,10]. The *GSL-AOP* locus was shown to occur in three alternative allelic states that give rise to three different leaf GSL profiles (*AOP2*, *AOP3* and *AOPnull*) (Fig. 1a) [11,12].

Interestingly, *AOP* genes are not only enzyme-coding and regulators of the glucosinolate pathway[13–17], but also exhibit regulatory roles in jasmonate signaling[13], nitrogen starvation adaptation[18], the circadian clock[19], and the onset of flowering[17,20]. Compared to other polymorph GSL loci, *GSL-AOP* shows the strongest contribution to total aliphatic GSL levels in an environment dependent manner[21]. With the aim of illuminating the structure-phenotype relationship of this multi-functional locus in comparison to other loci involved in GSL metabolism and flowering time control, we combined flowering time analysis and GSL phenotyping across 74 natural *A. thaliana* accessions, different tissues and conditions with targeted long read sequencing.

Results

ONT sequencing reveals novel structural variants of the *GSL-AOP* locus. We performed structure variance calling across the entire *GSL-AOP* locus using whole genome sequencing data available from the 1001 Genomes Consortium for 1135 *A. thaliana* accessions[22]. While the sequence diversity in the entire 1135 accession set was too high for SNP resolution clustering, it did allow for structure variance analysis caused by inversion and deletion events (InDels). The reference-guided genome assembly spanning a ~20kb region including all *AOP* and related genes was used for a locus specific analysis of the 1135 accessions. Using multiple sequence alignment and subsequent clustering by distance, we called haplotype groups according to structural differences resulting in four main haplotype clusters, H1 (22 accessions), the

most distant in the set, followed by H2 (26 accessions), H3 (45 accessions) and H4 (1042 accessions) (Extended Data Fig. 2). We extracted the frequencies of our haplotypes based on the ADMIXTURE classification[22] and found the most distinct group H1 to be geographically restricted to the Caucasus (Fig. 1b). Haplotype H1 is shared by seven populations in the Caucasus region represented by one accession from Iran (Anz-0), 16 from Azerbaijan (Istisu-1/5/9, Lerik1-3/4/7, Lerik2-1/3/6/7 and Nar-3/5) and five from Georgia (Lag2-2/4/6/7/10). Interestingly, the Lag2 population from Georgia has a sister population, Lag1-2/4/5/6/7/8, sampled from the same location but belonging to different sub clusters of haplotype H4. The Caucasus region thus potentially represents a geographical hotspot of local adaptation in the *GSL-AOP* locus evident by multiple genetic rearrangements.

To test if the structural variance in the *GSL-AOP* locus in the Caucasian accessions is reflected in the GSL profiles, we selected a set of 74 accessions including the entire H1 clade, their geographically closest accessions from other haplotype groups, and accessions representing the remaining clusters and geographical distribution of *A. thaliana* for detailed genetic and phenotypic analyses (Extended Data Tab. 1). Re-alignment of the *GSL-AOP* sequences from the selected accessions revealed two deletions in the H1 accessions spanning approximately 1.5 and 6 kb in the *GSL-AOP* locus (Extended Data Fig. 3). Interestingly, the *AOP3* promoter and most parts of the *AOP2* gene seemed to be missing in haplotype H1. Many other accessions exhibited additional, smaller deletions compared to the reference genome Col-0, especially in the intergenic region between *AOP3* and *AOP2*.

Despite the effort taken to access the genetic diversity within the *A. thaliana* genome, InDel or repeat polymorphisms larger then ~30bp are not accurately resolved by reference-guided assembly and short-read sequencing[22–28]. We therefore used targeted ONT amplicon sequencing to further investigate the structural variation in the *GSL-AOP* locus across our set of accessions. In addition to the *GSL-AOP* locus, we included the loci *GSL-OH* and *ESP*, responsible for the side chain modification and involved in metabolization of *AOP2*-modified GSLs, as well as the flowering loci *FLC*, *FRI*, *FT* and *VRN1*. While we found *FRI*, *FT* and *VRN1* to only harbor a few individual accessions with pronounced InDel events, *ESP*, *FLC* and *GSL-OH* are characterized by higher diversity (Extended Data Fig. 4). Variation in the *GSL-AOP* locus is associated with qualitative differences in GSL profiles, as the presence or absence of AOP enzymes in a given tissue results in absence or presence of their GSL products. Similarly, variation in *GSL-ELONG* and *ESP* is linked to side GSL chain length and differential GSL activation, respectively. All three loci display a complex architecture that has been shaped by independent structural events, suggesting structural variation as a common feature of adaptive loci with bimodal qualitative variation.

The ONT data allowed for rigorous clustering of the *GSL-AOP* locus into distinct groups with less pronounced between-group differences compared to 1001Genomes data illustrating the improved resolution of structural diversity by long-read sequencing. A syntenic block window adjustment of a gap size >40bp was applied to detect medium to large structure variance between the accessions. Within the *GSL-AOP* locus, several kb spanning rearrangements were detected. Cluster 6, former H1 accessions, displays 1) a deletion of *AOP2* and the intergenic region spanning towards *AOP3*, 2) an inversion and swap of *AOP3* (excluding UTRs) to the *AOP2* promoter, 3) a deletion of the sequence between *AOP3* and *At4g03038* (unknown function), including most of At4g03040 (unknown function) annotated in this region, and 4) a 1-kb insertion upstream of *MIR826a*/*b* (At4g03039/At4g04365), known to target the *AOP2* transcript (Fig. 2 (Lag2-2) and Extended Data Fig. 5)[18,29]. This haplotype is locally restricted to the Caucasian accessions Anz-0, Istisu, Lag2,

Nar and Xan, whereas the sister population Lag1 instead harbors an allele similar to the reference sequence Col-0 (cluster 4). It must be noted that locus-wide classification as performed here is not sensitive to deleterious point or frameshift mutations. As an example, this approach does not differentiate between Col-0 (*AOPnull* GSL profile) and Cvi-0 (*AOP2*). Two new distant alleles characterized by a ~2.7-kb and a ~6.2-kb intergenic insertion were identified in Agu-1 (cluster 3), Nok-3 (cluster 3) and in Spro-1 (cluster 2), respectively (Fig. 2). In these three accessions, the enzymatic function of *AOP2* was retained. Accession T480 (cluster 4) displays only two smaller deletions and one insertion up- and downstream of *AOP2* as compared to Cvi-0 and also shows Cvi-like GSL profiles (Fig. 2 and Fig. 3).

For seven accessions, we obtained an additional, shorter amplicon (Fig. 2), either amplified from a second *GSL-AOP* locus in those genomes or possibly indicating heterozygocity. More likely, these sequences represent adjacent parts in the *GSL-AOP* locus, assuming that the architecture of these seven accessions is similar to that in L*er*, Eri and Kyo, recently de-novo genome assembled by long read sequencing[5] (Fig. 2). In contrast to earlier suggestions[11], these accessions possess two copies of *AOP2* (*AOP2-1* and *AOP2-2*) and three copies of *AOP3* (*AOP3-1, AOP3-2* and *AOP3-3*), more than doubling the locus size. The strong structural deviation as evident from recently published long-read sequencing data[5] might explain why obtaining reliable sequences from this accession L*er* and 26 other accessions was not possible with targeted amplicon sequencing. We did, however, find six accessions with an additional *AOP2* copy (Lag1-2, Nie1-2, Pien, Agu-1, Rum-20, Pent-46) and one accession harboring an additional *AOP3* (Lag1-8), represented by the additional shorter amplicons. Although we also identified gene deletion events in *ESP* and *GSL-OH* (both in the accession Panik-1), the structural diversity in the *GSL-AOP* locus was not surpassed in the other tested loci (Fig. 1c and Fig. 2).

Across-tissue GSL profiling identifies new qualitative and quantitative chemotypes. We next analyzed the GLS profiles of the 74 selected accessions in leaves, roots, and seeds (Fig. 3A and Extended Data Fig. 6). Natural accessions have traditionally been classified into *AOP2*, *AOP3* and *AOPnull* accessions, dependent on their leaf GSL profile[12]. Leaves commonly accumulate predominantly AOP2- or less frequently AOP3-modified GSL, while the profile in seeds of the same plant can be the opposite or show accumulation of both AOP2- and AOP3-modified GLS. The Col-0 allele was named *AOPnull* and is used to describe plants accumulating only MT and MS GSL in leaves, although the presence of hydroxyalkyl GSL in the seeds indicates a functional *AOP3* in this accession[12].

Root and leaf tissue of the same genotype accumulated the same major class of SC GSL (Fig. 3A and Extended Data Fig. 6). Nevertheless, we found a number of accessions to co-accumulate AOP2- and AOP3-modified GSLs in either tissue. Roots of T480 contained low levels of AOP3-modified GSLs, which were not detectable in the leaves of the same genotype. Leaf-specific accumulation was observed for AOP2-modified GSLs in the Caucasus accessions (Lag2-2/6/10, Uk-3, Istisu-1/5/9, Nar-3), which lack an *AOP2* gene in the locus, suggesting the presence of an additional, so far unidentified gene encoding an enzyme with AOP2 activity. In contrast, accumulation of AOP3-modified GSLs in leaves but not roots (Lag1 accessions, UKSE06-252) indicates differential spatial expression patterns of *AOP3*, which could not have been predicted from the structure of the *GSL-AOP* locus.

In the seeds, all accessions accumulated AOP3-modified GSL structures[30] (Fig. 3A and Extended Data Fig. 6), which is congruent with the absence of an *AOP3* null allele in the ONT data set (Fig. 2). GSLs are not de novo synthesized in

A. thaliana seeds[31–33], but GSLs transported to developing seeds can be modified during seed maturation[34]. The presence of AOP-modified GSLs in seeds does therefore not reveal the tissue where they were synthesized. Nevertheless, most accessions that accumulated a high proportion of AOP3-modified GSLs in seeds, also contained these GSLs in leaf and root tissue. This was particularly pronounced for the Caucasus accessions lacking an *AOP2* gene in the locus (cluster 6) (Fig. 3A and Extended Data Fig. 6). Leaves and roots of Löv-5 and Eden-1 contained almost exclusively AOP2-modified GSLs, but interestingly, these structures were absent from the seeds in both accessions, which is likely due to structure-specific seed loading of GSLs. The accession Rev-1 did not accumulate detectable levels of either AOP2- or AOP3-modified GSLs in leaves and roots, however, both GSL classes were found in seeds, suggesting co-expression of *AOP2* and *AOP3* in developing seeds. Löv-5, Eden-1 and Rev-1 display very diverse GSL phenotypes, but we were unable to obtain *GSL-AOP* ONT data for those accessions possibly indicating that they carry structurally different *GSL-AOP* loci than those covered by our approach.

Seed GSL levels (min-max: 1.2-3.8 nmol/mg) were more uniform across all tested accessions compared to root GSL (min-max: 0.6-18.2 nmol/mg) and leaf GSL (min-max: 0.5-7.1 nmol/mg). The relative GSL accumulation in the different tissues of one genotype varied largely, i.e. accessions with low levels in leaves and roots accumulated seed GSL levels above average (e.g. Uk-3) (Extended Data Tab. 2). Levels of total SC GSLs in leaves were generally not mirrored in roots. The levels of AOP2- and AOP3-modified GSLs in leaves has been reported to positively correlate with total leaf SC GSL[12,14,15,17,35]. In agreement with these earlier findings, accessions which lack an enzymatically functional *AOP2* in the *GSL-AOP* locus accumulated overall lower levels of SC GSLs in leaves (Fig. 3B-C). In contrast, the opposite was observed for seeds, i.e. the total SC GSL levels were lower in seeds of AOP2null accessions. Taken together, GSL analysis of leaves, roots, and seeds showed that overall distribution patterns of AOP-modified structures can largely, but not entirely, be predicted by the structure of the *GSL-AOP* locus. The tissue-specific GSL profiles pointed to either natural variation in long-distance GSL transport or tissue-specific differences in AOP expression.

The modular organization of the *GSL-AOP* locus is attained through variation in *AOP* gene copy number and orientation, accompanied by smaller and larger InDels in the intergenic regions (Fig. 2). The cluster 6 accessions (e.g. Lag2) from the Caucasus region further display an inversion, which renders *AOP3* to be controlled by the promoter that leads to strong *AOP2* expression in Cvi-0. We aligned ~1Kb of the putative *AOP3* promoter sequences and found them to cluster in three major groups, partially overlapping with the major *GSL-AOP* locus clusters. The *AOP3* promoters from most Caucasus accessions (Lag2, Lerik, Istisu, Xan) grouped together in promoter cluster 3 (Fig. 4A). These accessions are characterized by the absence of AOP2-modified GSLs in all tissues analyzed (Fig. 4B) and by high levels of AOP3-modified GSL; up to 10-fold and 3-fold higher than in L*er* in leaves and roots, respectively (Fig. 4C).

L*er* harbors three *AOP3* genes (Fig. 2), all comprising an intact *AOP3* ORF. The promoter of *AOP3-1* is highly similar to the Col-0 *AOP3* promoter sequence (cluster 1), whereas the sequences of L*er AOP3-2* and *AOP3-3* (cluster 2) are more similar and reverse complement to the *AOP3* promoter sequences of cluster 3. Despite the three apparently functional *AOP3* gene copies, L*er* accumulates considerably lower levels of AOP3-modified GSL, indicating that there is no gene dose effect on the GSL phenotype. Overall, our findings illustrate that gene duplications, inversions, and promoter swapping within the *GSL-AOP* locus, which had not been revealed by short-read sequencing, contribute to the qualitative and quantitative variation in tissue-specific GSL profiles.

The large genetic variation in the Caucasus region is not restricted to the *GSL-AOP* locus. ONT sequencing of the *GSL-AOP* locus revealed novel haplotypes in the Caucasus region. Clustering according to the overall locus architecture identified a distinct haplotype in Azerbaijan (Istisu-1/5/9, Lerik1-3/4/7, Lerik2-1/3/6/7 and Nar-3/5), Iran (Anz-0) and Georgia (Lag2-4/6/7/10). These accessions, which appear to lack an *AOP2* gene in the locus, accumulate high levels of AOP3-modified GSLs with a C3 side chain across all analyzed tissues (Extended Data Tab. 2). The Lag1 accessions, also collected in Georgia, belong to the Col-0 like haplotype and produce mainly AOP2-modfied GSLs with variable C3:C4 side chain ratios. In the *ESP* locus, earlier shown to act epistatically to *GSL-AOP*[36], no larger sequence differences were detected among the Caucasian accessions. The presence of a functional AOP2 enzyme and its C4 precursor in Lag1 accessions leads to accumulation of 3-butenyl GSL, which can be hydroxylated by GSL-OH, whereas this GSL-OH substrate is absent in the accessions of the Lag2 sister clade (Extended Data Tab. 2). Nevertheless, Lag1 and Lag2 share the same *GSL-OH* haplotype, which is distinct from that of the other Caucasian accessions. Genetic variation in the *GSL-OH* locus in this geographical region thus does not correlate with the presence of the known GSL-OH enzyme substrate.

Next, we set out to investigate whether the regional genetic variation in *GSL-AOP* and *GSL-OH* is limited to GSL-related loci and expanded our analysis to loci involved in flowering time control. Previous studies have linked the *GSL-AOP* to the circadian clock[19], expression of the major flowering time repressor *FLC*[37], and the onset of flowering[20]. Similar to *GSL-AOP* and *GSL-OH*, two *FLC* haplotypes were found in the Caucasian accessions, representing the two major *FLC* clusters in our set of accessions. Whereas the *FLC* locus in the Lag1 and Lag2 accessions shows high similarity to that in the reference accession Col-0, the *FLC* promoter region in the accessions from Azerbaijan is characterized by several InDels (Fig. 5 and Extended Data Fig. 7). No similar geographical differences were seen for *FRI*, *FT* and *VRN1*.

We phenotyped our set of natural accessions for flowering time in inductive long day (LD) conditions and after vernalization to assess potential differences in the phenotypic responses that may hint to local adaption in the Caucasus region (Fig. 6, Extended Data Tab. 3). We found a continuous distribution for flowering time under LD conditions spanning from early flowering accessions like UKID96 or L*er*-0 at ~31 days to accessions not flowering within a 200-day-growing period, e.g. Löv-5 and Agu-1. All accessions responded to the vernalization treatment, reducing the median flowering time from 102 days in LD to 30 days after vernalization. The accessions from Georgia (Lag1, Lag2) with the exception of Lag1-2 flowered after 80-100 days under LD conditions and those from Azerbaijan (Istisu-1/5/9, Lerik1-3/4/7, Lerik2-1/3/6/7 and Nar-3/5) even later (>100 days). All showed a pronounced vernalization response. When comparing flowering time across the two conditions, the Caucasian accessions cluster according to their *FLC* (cluster 5) and *GSL-AOP* (cluster 6) haplotypes (Fig. 6C).

The levels of total SC GSLs in leaves, representing the most abundant GSL class in this tissue, ranged from 0.07 to 6.1 nmol/mg FW in plants cultivated without vernalization and from 0.06 to 21 nmol/mg FW in vernalized plants (Fig. 6B). We did not observe any qualitative changes in leaf GSL profiles (Extended Data Tab. 2). Lag1-2/4/6/7 showed a five-fold difference in SC GSL accumulation reaching ca. 10 nmol/mg FW after vernalization. Lag1-8 and Lag1-5 displayed higher levels under LD conditions, but similar levels in response to vernalization. In contrast, all other Caucasian accessions showed only a two-fold change and grouped tightly when plotted according to SC GSL levels. We found *GSL-AOP* as well as *FLC* to show higher variation in locus structure among accessions collected in the Caucasus region than the other

studied loci. The observation that accessions with different GSL profiles differed in their flowering time response to vernalization further supports the functional connection between these two loci.

Discussion

Using a targeted long-read sequencing approach, we demonstrated that the *GSL-AOP* locus underwent recurrent local rearrangement, contributing to the chemical diversity in the Caucasus region. The locus represents a hotspot for rearrangements that accumulated different types of mutations, which may enable rapid responses to changes in the biotic environment[5]. Which AOP-modified GSLs accumulate in an accession depends on the allelic status of *GSL-ELONG*; and their chemical diversity is amplified by variation in *GSL-OH* and *ESP*. All four loci are structurally variable (Fig. 2, Extended Data Fig. 4 and [38]) indicating that structural events in loci with bimodal phenotypic output provide a molecular mechanism for local adaption. By phenotyping a diverse panel of *A. thaliana* accessions, we found tissue-specific qualitative and quantitative variation in GSL profiles, which can be explained by presence and absence of *AOP* genes in the *GSL-AOP* locus, tissue-specific *AOP* expression and likely also natural variation in GSL long-distance transport. Differences between root and leaf GSLs illustrate different regulatory networks, which enable plants to cope with tissue-specific pests attacking specific tissues e.g. nematodes, gnats and clubroot disease[39–41]. Similar to the *GSL-AOP* locus, *FLC* has been a target of repeated rearrangements in *A. thaliana* and other Brassicaceae [42–48]; and like *GSL-AOP*, the *FLC* locus showed increased genetic variation in the Caucasus region.

ONT sequencing of the *GSL-AOP* locus revealed a novel allele missing the *AOP2* gene, due to an inversion deletion event, in some Caucasian accessions. Other *A. thaliana* accessions including Col-0 represent loss-of-function alleles due to frame shift mutations and loss of the entire gene, as recently also reported for *Erysimum cheiranthoides* [49], indicate that *AOP2* is expendable. In contrast, no *AOP3* loss-of-function mutant has been reported to date, which might be linked to specific defensive properties of AOP3-modified GSL present in the seeds and thus the seedlings of all *A. thaliana* genotypes. Constitutive expression of *AOP3* in leaves does not occur in all accessions but can be induced upon attack[50]. The AOP3 product 3-hydroxypropyl (3-OHP) GSL provides regulatory input to the ancient TOR pathway, which balances growth and development with the availability of nutrients and energy. This remarkable signaling function of 3-OHP GSL or derived metabolites raises the additional question whether the positive selection pressure on *AOP3* across environments is related to defense or/and development.

Material & Methods

Plant lines and growth conditions

Arabidopsis thaliana accessions were received from The Nottingham Arabidopsis Stock Center (for additional information see Extended Data Table 1). Plants were cultivated in growth chambers under long day conditions 16h (LD)

21°C day/18°C night and 60% relative humidity. Light intensity was adjusted to ~150µE/(m2*s). Plants used for phenotyping were grown on a soil (Pindstrup nr. 2, Pindstrup Mosebrug A/S, Denmark) sand mixture (3:1) and received four days stratification at 4°C in darkness, before being moved to the growth chamber. Plants were sown in a randomized block design. Days to bolting were scored when the flowering bolt reached 1 cm in height. Days to flowering were scored when first flower but was fully expanded.

Plants used for vernalization studies were first grown on plates with half strength Murashige and Skoog medium (Duchefa Biochemie, M0222) and 0.8% agar (Sigma, A9799). Plates received four days stratification at 4°C in darkness, before being placed in the growth chamber under short day conditions 8h (SD) 21°C day/18°C night and 60% relative humidity. Light intensity was adjusted to ~150µE/(m2*s). After one week, plates were transferred to the cold room at 4°C and 8h (SD) light (~40µE/(m2*s) for six weeks. After vernalization treatment, plants were transplanted to soil and placed in growth chambers under long day conditions 16h (LD), ~150µE/(m2*s), 21°C day/18°C night and 60% relative humidity until flowering. Experiments were stopped after 200 days (LD) and 130 days (LD after vernalization); plants that had not flowered until end of the experiment were scored as flowering on the last day of experiment.

Analysis of glucosinolate profiles

Plants from the LD flowering time experiment were used for GLS analysis. From 3-week-old plants, a single rosette leaf (leaf number six) was harvested and weighed, roots from the same plants were rinsed with water to remove remaining soil and were subsequently weighed (n=3). Seed analysis was performed on single seeds (n=3). Plants from the vernalization experiment were used for leaf GLS analysis. Three weeks after vernalization, a single rosette leaf (leaf number six) was harvested. GSLs were analyzed and quantified as desulfo-GSLs by LC/triple quadrupole-MS (alternate protocol 2[51]), using 3 nmol *p*-OH-benzyl GSL per sample as an internal standard. A list of the GSLs covered by the analysis can be found in Extended Data Table 2. Calculation for chemotype model: For each tissue we calculated percent alkenyl, hydroxyalkyl and MS of total SC.

High molecular weight DNA isolation

5-week-old plants were kept in darkness for 3 days to reduce starch and anthocyanin accumulation. Leaves from single plants were harvested in 15-ml tubes containing 3-mm bearing balls, frozen in liquid nitrogen and homogenized by shaking in a paint shaker (2 x 3 min at max. speed). Frozen leaf powder was extracted with 5ml CTAB lysis buffer (2% CTAB, 100mM TRIS-HCl, 20mM EDTA, 1.4 M NaCl, pH 8.0) containing 0.2% 2-mercaptoethanol and 0.5µl RNase A/T1 (ThermoScientific) for 1 hour at 65°C. DNA was purified by adding 1 volume chloroform, mixed by inversion and centrifuged for 30min at 5000rpm/4°C. Supernatant was mixed with 0.75 volume ice cold Isopropanol for DNA precipitation at -80°C for 25min. Followed by centrifugation for 30 min at 5000 rpm/4°C. The DNA pellet was washed twice in 5 ml ice-cold 70% (v/v) ethanol, dried, resuspended in 500 µl H2O overnight and stored at 4°C. Presence of high molecular weight DNA was verified on 0.6% agarose gels and DNA quality and concentration were evaluated by NanoDrop ND-1000 spectrophotometer (Saveen and Werner AB, Sweden).

Long range PCR, barcoding and library preparation

PCR1 for amplification of loci of interest was performed using primers with 5' overhang designed to incorporate ONT compatible adapters in 2nd PCR as followed: 2µl DNA(~100ng), 1µl of corresponding primers (10µM each), 2µl dNTP mixture (2.5mM each), 0.2µl PrimeSTAR® GXL DNA Polymerase (Takara), 9.8µl nuclease free H20. PCR conditions: Denaturation for 1min at 98°C followed by 30 cycles of 10s at 98°C, 20s at 55°C and 15min at 68°C, and final elongation for 15min at 68°C. Amplicons were verified on 0.6% agarose gels.

Amplicons from each of sample were barcoded using a set of 96 ONT compatible barcodes. The barcoding oligonucleotides are modified from previously reported adapters (Extended Data Tab. 4). Each oligo was composed of a 15-bp spacer to ensure higher tolerance for the low-quality at the beginning of the DNA strand entering the pore and thus higher recovery of barcode sequence following the ONT compatible barcode sequence and the lastly PCR1 complementary region (Extended Data Tab. 4). The PCR2 reaction mix contained 2 µl PCR1, 1 µl of corresponding primer (10 µM each), 2 µl dNTP mixture (2.5 mM each), 0.2 µl PrimeSTAR® GXL DNA Polymerase (Takara), 9.8 µl nuclease free H20. PCR temperature program: Denaturation for 1 min at 98°C followed by 30 cycles of 10 s at 98°C, 20 s at 55°C and 15 min at 68°C, and final elongation for 15 min at 68°C.

Barcoded amplicons were verified on 0.6% agarose gels and pooled at equimolar ratios. The pooled library was cleaned with AMPure XP beads (Beckman Coulter Genomic, CA, USA) in a 1:1 volume. The bead pellet was washed with 80% (v/v) ethanol and re-suspended in 100 µl of nuclease-free water. The sequencing library was prepared using 1D Amplicon by ligation Kit (SQK-LSK109) according to the manual (version: GDE_9063_v109_revA_23May2018). ~0.5 µg of amplicons were used for the initial step of end-prep. ~150 ng of prepared amplicon library were loaded on an R9.4.1 flow cell.

Nanopore amplicon sequencing data analysis

Raw data were collected using Oxford Nanopore software: MinKnow 19.06.8 (https://nanoporetech.com). Guppy 3.2.2 basecalling toolkit was used to base call raw fast5 to fastq (https://nanoporetech.com). Porechop v0.2.2 was used for adapter trimming and sample demultiplexing (https://github.com/rrwick/Porechop). Porechop settings adapter list was (adapters.py).

Sequences containing quality scores (fastq files) were quality corrected using NanoFilt (q ≥ 10; read length > 1Kb)[52]. Subsequently, reads within a respective length range were extracted with Cutadapt v1.15[53], and corrected with Canu v1.6[54] using the following parameters: genomeSize=5k, minimumReadLength=200, correctedErrorRate=0.05, corOutCoverage=5000, corMinCoverage=2 and minOverlapLength=50. The corrected reads were sorted-by-length and clustered with cluster_fast from VSEARCH[55], using the following options: -id of 0.9, -minsl of 0.8, -sizeout, and min_cons_pct of 20. Corrected reads were sorted by alignment against loci reference[56] and manually verified in CLC Main Workbench 7.9.1. Reads not containing the flanking sequence of PCR1 primer were rejected.

Data analysis

R version 3.5.1 (2018-07-02) "Feather Spray" was used for data analysis[57]. Data processing was conducted by the use of plyr[58], reshape2[59] and tidyr[60]. Geographical distribution was visualized by ggmap[61] and further plots were conducted by the use of ggplot2[62].

1001Genome and ONT data analysis

Desired sequences were downloaded from (https://tools.1001genomes.org/pseudogenomes) and processed using DECIPHER v2.0[56]. For the alignments, gaps were removed and multiple sequence alignments were performed using the following settings: iterations=15, refinements=15, gapOpening=c(-30, -14), gapExtension=c(-4,-1). Similarity clustering was performed on distance matrix using Neighbor-Joining method and adjusted cutoff.

Syntenic blocks for ONT data were identified using FindSynthey[56] with the following modified settings: maxSep=40, maxGap=40. Sequences are orientated according to Col-0 reference sequence. Dot- and syntenic block plots are shown for representative accessions.

References

1. Fournier-Level, A., Korte, A., Cooper, M. D., Nordborg, M., Schmitt, J. & Wilczek, A. M. A map of local adaptation in Arabidopsis thaliana. *Science* **334,** 86–89 (2011).

2. Ristova, D., Giovannetti, M., Metesch, K. & Busch, W. Natural genetic variation shapes root system responses to phytohormones in Arabidopsis. *Plant J.* **96,** 468–481 (2018).

3. Gaudinier, A., Rodriguez-Medina, J., Zhang, L., Olson, A., Liseron-Monfils, C., Bågman, A.-M., Foret, J., Abbitt, S., Tang, M., Li, B., Runcie, D. E., Kliebenstein, D. J., Shen, B., Frank, M. J., Ware, D. & Brady, S. M. Transcriptional regulation of nitrogen-associated metabolism and growth. *Nature* **563,** 259–264 (2018).

4. Zhang, W., Corwin, J. A., Copeland, D., Feusier, J., Eshbaugh, R., Chen, F., Atwell, S. & Kliebenstein, D. J. Plastic Transcriptomes Stabilize Immunity to Pathogen Diversity: The Jasmonic Acid and Salicylic Acid Networks within the Arabidopsis/Botrytis Pathosystem. *Plant Cell* **29,** 2727–2752 (2017).

5. Jiao, W.-B. & Schneeberger, K. Chromosome-level assemblies of multiple Arabidopsis genomes reveal hotspots of rearrangements with altered evolutionary dynamics. *Nat. Commun.* **11,** 989 (2020).

6. Wittstock, U. & Burow, M. Glucosinolate breakdown in Arabidopsis: mechanism, regulation and biological significance. *Arabidopsis Book* **8,** e0134 (2010).

7. Kroymann, J., Textor, S., Tokuhisa, J. G., Falk, K. L., Bartram, S., Gershenzon, J. & Mitchell-Olds, T. A gene controlling variation in Arabidopsis glucosinolate composition is part of the methionine chain elongation pathway. *Plant Physiol.* **127,** 1077–1088 (2001).

8. Li, J., Hansen, B. G., Ober, J. A., Kliebenstein, D. J. & Halkier, B. A. Subclade of flavin-monooxygenases involved in aliphatic glucosinolate biosynthesis. *Plant Physiol.* **148,** 1721–1733 (2008).

9. Hansen, B. G., Kliebenstein, D. J. & Halkier, B. A. Identification of a flavin-monooxygenase as the S-oxygenating enzyme in aliphatic glucosinolate biosynthesis in Arabidopsis. *Plant J.* **50,** 902–910 (2007).

10. Ibdah, M. & Pichersky, E. Arabidopsis Chy1 null mutants are deficient in benzoic acid-containing glucosinolates in the seeds. *Plant Biol. (Stuttg.)* **11,** 574–581 (2009).

11. Chan, E. K. F., Rowe, H. C. & Kliebenstein, D. J. Understanding the evolution of defense metabolites in Arabidopsis thaliana using genome-wide association mapping. *Genetics* **185,** 991–1007 (2010).

12. Kliebenstein, D. J., Lambrix, V. M., Reichelt, M., Gershenzon, J. & Mitchell-Olds, T. Gene duplication in the diversification of secondary metabolism: tandem 2-oxoglutarate-dependent dioxygenases control glucosinolate biosynthesis in Arabidopsis. *Plant Cell* **13,** 681–693 (2001).

13. Burow, M., Atwell, S., Francisco, M., Kerwin, R. E., Halkier, B. A. & Kliebenstein, D. J. The Glucosinolate Biosynthetic Gene AOP2 Mediates Feed-back Regulation of Jasmonic Acid Signaling in Arabidopsis. *Mol. Plant* **8,** 1201–1212 (2015).

14. Wentzell, A. M., Rowe, H. C., Hansen, B. G., Ticconi, C., Halkier, B. A. & Kliebenstein, D. J. Linking metabolic QTLs with network and cis-eQTLs controlling biosynthetic pathways. *PLoS Genet.* **3,** 1687–1701 (2007).

15. Li, G. & Quiros, C. F. In planta side-chain glucosinolate modification in Arabidopsis by introduction of dioxygenase Brassica homolog BoGSL-ALK. *Theor. Appl. Genet.* **106,** 1116–1121 (2003).

16. Rohr, F., Ulrichs, C., Schreiner, M., Nguyen, C. N. & Mewis, I. Impact of hydroxylated and non-hydroxylated aliphatic glucosinolates in Arabidopsis thaliana crosses on plant resistance against a generalist and a specialist herbivore. *Chemoecology* **21,** 171–180 (2011).

17. Jensen, L. M., Kliebenstein, D. J. & Burow, M. Investigation of the multifunctional gene AOP3 expands the regulatory network fine-tuning glucosinolate production in Arabidopsis. *Front. Plant Sci.* **6,** 762 (2015).

18. He, H., Liang, G., Li, Y., Wang, F. & Yu, D. Two young MicroRNAs originating from target duplication mediate nitrogen starvation adaptation via regulation of glucosinolate synthesis in Arabidopsis thaliana. *Plant Physiol.* **164,** 853–865 (2014).

19. Kerwin, R. E., Jimenez-Gomez, J. M., Fulop, D., Harmer, S. L., Maloof, J. N. & Kliebenstein, D. J. Network quantitative trait loci mapping of circadian clock outputs identifies metabolic pathway-to-clock linkages in Arabidopsis. *Plant Cell* **23,** 471–485 (2011).

20. Jensen, L. M., Jepsen, H. S. K., Halkier, B. A., Kliebenstein, D. J. & Burow, M. Natural variation in cross-talk between glucosinolates and onset of flowering in Arabidopsis. *Front. Plant Sci.* **6,** 697 (2015).

21. Kerwin, R., Feusier, J., Corwin, J., Rubin, M., Lin, C., Muok, A., Larson, B., Li, B., Joseph, B., Francisco, M., Copeland, D., Weinig, C. & Kliebenstein, D. J. Natural genetic variation in Arabidopsis thaliana defense metabolism genes modulates field fitness. *Elife* **4,** (2015).

22. The 1001 Genomes Consortium. 1,135 Genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. *Cell* **166,** 481–491 (2016).

23. Ossowski, S., Schneeberger, K., Clark, R. M., Lanz, C., Warthmann, N. & Weigel, D. Sequencing of natural strains of Arabidopsis thaliana with short reads. *Genome Res.* **18,** 2024–2033 (2008).

24. Clark, R. M., Schweikert, G., Toomajian, C., Ossowski, S., Zeller, G., Shinn, P., Warthmann, N., Hu, T. T., Fu, G., Hinds, D. A., Chen, H., Frazer, K. A., Huson, D. H., Schölkopf, B., Nordborg, M., Rätsch, G., Ecker, J. R. & Weigel, D. Common sequence polymorphisms shaping genetic diversity in Arabidopsis thaliana. *Science* **317,** 338–342 (2007).

25. Cao, J., Schneeberger, K., Ossowski, S., Günther, T., Bender, S., Fitz, J., Koenig, D., Lanz, C., Stegle, O., Lippert, C., Wang, X., Ott, F., Müller, J., Alonso-Blanco, C., Borgwardt, K., Schmid, K. J. & Weigel, D. Whole-genome sequencing of multiple Arabidopsis thaliana populations. *Nat. Genet.* **43,** 956–963 (2011).

26. Long, Q., Rabanal, F. A., Meng, D., Huber, C. D., Farlow, A., Platzer, A., Zhang, Q., Vilhjálmsson, B. J., Korte, A., Nizhynska, V., Voronin, V., Korte, P., Sedman, L., Mandáková, T., Lysak, M. A., Seren, Ü., Hellmann, I. & Nordborg, M. Massive genomic variation and strong selection in Arabidopsis thaliana lines from Sweden. *Nat. Genet.* **45,** 884–890 (2013).

27. Gan, X., Stegle, O., Behr, J., Steffen, J. G., Drewe, P., Hildebrand, K. L., Lyngsoe, R., Schultheiss, S. J., Osborne, E. J., Sreedharan, V. T., Kahles, A., Bohnert, R., Jean, G., Derwent, P., Kersey, P., Belfield, E. J., Harberd, N. P., Kemen, E., Toomajian, C., Kover, P. X., Clark, R. M., Rätsch, G. & Mott, R. Multiple reference genomes and transcriptomes for Arabidopsis thaliana. *Nature* **477,** 419–423 (2011).

28. Schneeberger, K., Ossowski, S., Ott, F., Klein, J. D., Wang, X., Lanz, C., Smith, L. M., Cao, J., Fitz, J., Warthmann, N., Henz, S. R., Huson, D. H. & Weigel, D. Reference-guided assembly of four diverse Arabidopsis thaliana genomes. *Proc. Natl. Acad. Sci. USA* **108,** 10249–10254 (2011).

29. Liang, G., He, H. & Yu, D. Identification of nitrogen starvation-responsive microRNAs in Arabidopsis thaliana. *PLoS One* **7,** e48951 (2012).

30. Kliebenstein, D. J., Kroymann, J., Brown, P., Figuth, A., Pedersen, D., Gershenzon, J. & Mitchell-Olds, T. Genetic control of natural variation in Arabidopsis glucosinolate accumulation. *Plant Physiol.* **126,** 811–825 (2001).

31. Field, B., Cardon, G., Traka, M., Botterman, J., Vancanneyt, G. & Mithen, R. Glucosinolate and amino acid biosynthesis in Arabidopsis. *Plant Physiol.* **135,** 828–839 (2004).

32. Haughn, G. W., Davin, L., Giblin, M. & Underhill, E. W. Biochemical Genetics of Plant Secondary Metabolites in Arabidopsis thaliana: The Glucosinolates. *Plant Physiol.* **97,** 217–226 (1991).

33. Nour-Eldin, H. H., Andersen, T. G., Burow, M., Madsen, S. R., Jørgensen, M. E., Olsen, C. E., Dreyer, I., Hedrich, R., Geiger, D. & Halkier, B. A. NRT/PTR transporters are essential for translocation of glucosinolate defence compounds to seeds. *Nature* **488,** 531–534 (2012).

34. Kliebenstein, D. J., D'Auria, J. C., Behere, A. S., Kim, J. H., Gunderson, K. L., Breen, J. N., Lee, G., Gershenzon, J., Last, R. L. & Jander, G. Characterization of seed-specific benzoyloxyglucosinolate mutations in Arabidopsis thaliana. *Plant J.* **51,** 1062–1076 (2007).

35. Burow, M., Halkier, B. A. & Kliebenstein, D. J. Regulatory networks of glucosinolates shape Arabidopsis thaliana fitness. *Curr. Opin. Plant Biol.* **13,** 348–353 (2010).

36. Wentzell, A. M., Boeye, I., Zhang, Z. & Kliebenstein, D. J. Genetic networks controlling structural outcome of glucosinolate activation across development. *PLoS Genet.* **4,** e1000234 (2008).

37. Atwell, S., Huang, Y. S., Vilhjálmsson, B. J., Willems, G., Horton, M., Li, Y., Meng, D., Platt, A., Tarone, A. M., Hu, T. T., Jiang, R., Muliyati, N. W., Zhang, X., Amer, M. A., Baxter, I., Brachi, B., Chory, J., Dean, C., Debieu, M., de Meaux, J., Ecker, J. R., Faure, N., Kniskern, J. M., Jones, J. D. G., Michael, T., Nemri, A., Roux, F., Salt, D. E., Tang, C., Todesco, M., Traw, M. B., Weigel, D., Marjoram, P., Borevitz, J. O., Bergelson, J. & Nordborg, M. Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. *Nature* **465,** 627–631 (2010).

38. Kroymann, J., Donnerhacke, S., Schnabelrauch, D. & Mitchell-Olds, T. Evolutionary dynamics of an Arabidopsis insect resistance quantitative trait locus. *Proc. Natl. Acad. Sci. USA* **100 Suppl 2,** 14587–14592 (2003).

39. van Dam, N. M., Tytgat, T. O. G. & Kirkegaard, J. A. Root and shoot glucosinolates: a comparison of their diversity, function and interactions in natural and managed ecosystems. *Phytochem. Rev.* **8,** 171–186 (2009).

40. Van Poecke, R. M. P. Arabidopsis-insect interactions. *Arabidopsis Book* **5,** e0107 (2007).

41. Tsunoda, T., Krosse, S. & van Dam, N. M. Root and shoot glucosinolate allocation patterns follow optimal defence allocation theory. *J Ecol* **105,** 1256–1266 (2017).

42. Schranz, M. E., Quijada, P., Sung, S.-B., Lukens, L., Amasino, R. & Osborn, T. C. Characterization and effects of the replicated flowering time gene FLC in Brassica rapa. *Genetics* **162,** 1457–1468 (2002).

43. Kemi, U., Niittyvuopio, A., Toivainen, T., Pasanen, A., Quilot-Turion, B., Holm, K., Lagercrantz, U., Savolainen, O. & Kuittinen, H. Role of vernalization and of duplicated FLOWERING LOCUS C in the perennial Arabidopsis lyrata. *New Phytol.* **197,** 323–335 (2013).

44. Wang, R., Farrona, S., Vincent, C., Joecker, A., Schoof, H., Turck, F., Alonso-Blanco, C., Coupland, G. & Albani, M. C. PEP1 regulates perennial flowering in Arabis alpina. *Nature* **459,** 423–427 (2009).

45. Li, P., Filiault, D., Box, M. S., Kerdaffrec, E., van Oosterhout, C., Wilczek, A. M., Schmitt, J., McMullan, M., Bergelson, J., Nordborg, M. & Dean, C. Multiple FLC haplotypes defined by independent cis-regulatory variation underpin life history diversity in Arabidopsis thaliana. *Genes Dev.* **28,** 1635–1640 (2014).

46. Lempe, J., Balasubramanian, S., Sureshkumar, S., Singh, A., Schmid, M. & Weigel, D. Diversity of flowering responses in wild Arabidopsis thaliana strains. *PLoS Genet.* **1,** 109–118 (2005).

47. Michaels, S. D., He, Y., Scortecci, K. C. & Amasino, R. M. Attenuation of FLOWERING LOCUS C activity as a mechanism for the evolution of summer-annual flowering behavior in Arabidopsis. *Proc. Natl. Acad. Sci. USA* **100,** 10102–10107 (2003).

48. Lee, C.-R., Hsieh, J.-W., Schranz, M. E. & Mitchell-Olds, T. The Functional Change and Deletion of FLC Homologs Contribute to the Evolution of Rapid Flowering in Boechera stricta. *Front. Plant Sci.* **9,** 1078 (2018).

49. Züst, T., Strickler, S. R., Powell, A. F., Mabry, M. E., An, H., Mirzaei, M., York, T., Holland, C. K., Kumar, P., Erb, M., Petschenka, G., Gomez, J.-M., Perfectti, F., Mueller, C., Pires, C., Mueller, L. A. & Jander, G. Rapid and independent evolution of ancestral and novel defenses in a genus of toxic plants (*Erysimum* , Brassicaceae). *BioRxiv* (2019). doi:10.1101/761569

50. Sato, Y., Tezuka, A., Kashima, M., Deguchi, A., Shimizu-Inatsugi, R., Yamazaki, M., Shimizu, K. K. & Nagano, A. J. Transcriptional Variation in Glucosinolate Biosynthetic Genes and Inducible Responses to Aphid Herbivory on Field-Grown Arabidopsis thaliana. *Front. Genet.* **10,** 787 (2019).

51. Crocoll, C., Halkier, B. A. & Burow, M. Analysis and quantification of glucosinolates. *Curr. Protoc. Plant Biol.* **1,** 385–409 (2016).

52. De Coster, W., D'Hert, S., Schultz, D. T., Cruts, M. & Van Broeckhoven, C. NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics* **34,** 2666–2669 (2018).

53. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet j.* **17,** 10 (2011).

54. Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H. & Phillippy, A. M. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* **27,** 722–736 (2017).

55. Rognes, T., Flouri, T., Nichols, B., Quince, C. & Mahé, F. VSEARCH: a versatile open source tool for metagenomics. *PeerJ* **4,** e2584 (2016).

56. Wright, E., S. Using DECIPHER v2.0 to Analyze Big Biological Sequence Data in R. *R J.* **8,** 352 (2016).

57. R: The R Project for Statistical Computing. at <https://www.R-project.org/.>

58. Wickham, H. The Split-Apply-Combine Strategy for Data Analysis. *J Stat Softw* **40,** (2011).

59. Wickham, H. Reshaping Data with the reshape Package. *J Stat Softw* **21,** (2007).

60. CRAN - Package tidyr. at <https://CRAN.R-project.org/package=tidyr>

61. David Kahle, Hadley Wickham and Scott Jackson. ggmap: Spatial Visualization with ggplot2.

62. Wickham, H. *ggplot2 - Elegant Graphics for Data Analysis* . (Springer-Verlag New York, 2016). doi:10.1007/978-0-387-98141-3

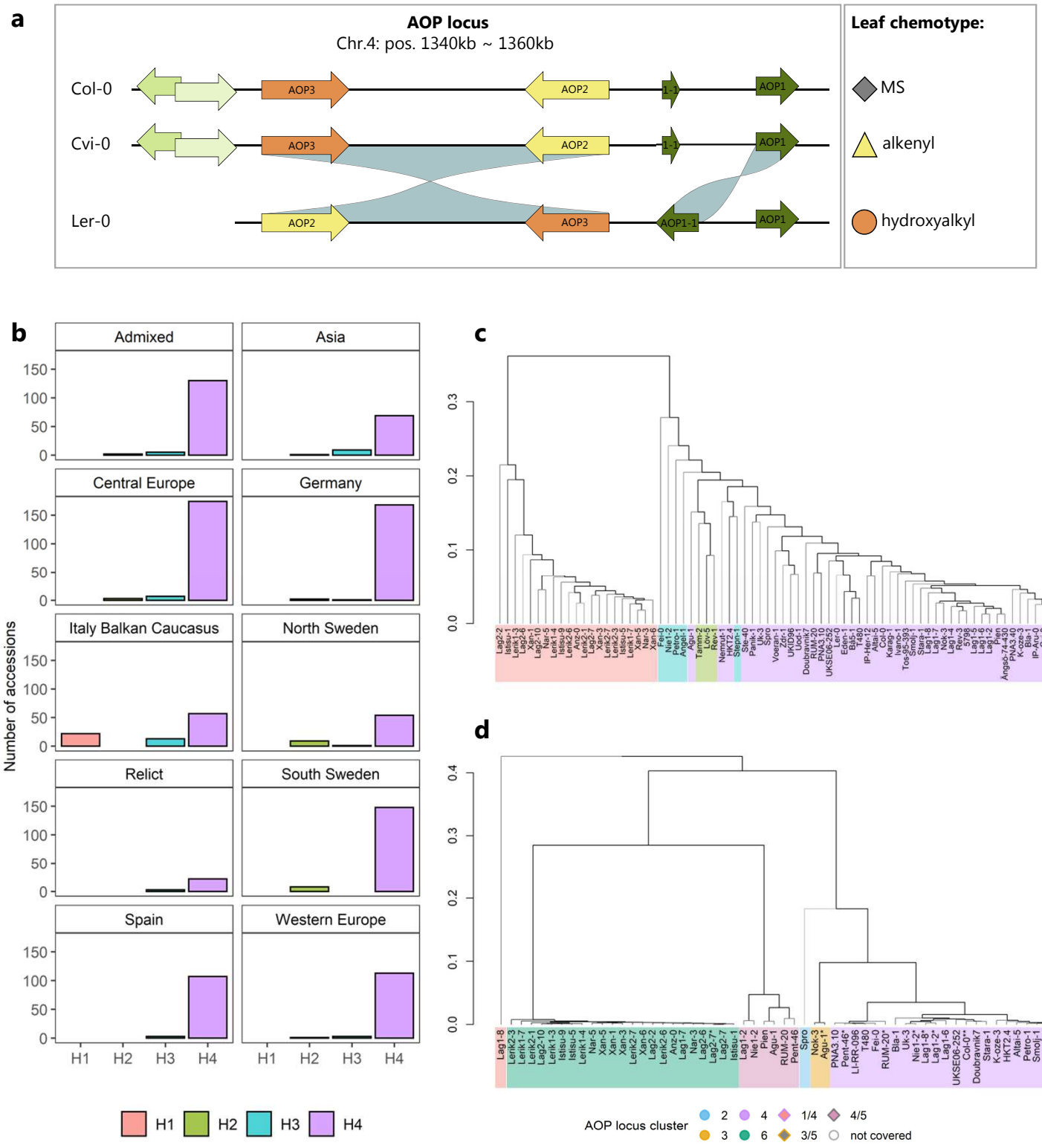

Figure 1 (see legend next page)

Figure 1

a *GSL-AOP* locus variation in *A. thaliana* determines the GSL leaf chemotype. Diagram of *GSL-AOP* locus structure in three accessions representing the main classical haplotype classes. Colored arrows show combinations of genes: *AOP1* (dark green, unknown function), *AOP2* (yellow, alkenyl producing), *AOP3* (orange, hydroxyalkyl producing) and flanking genes (green: At4g03038, unknown function and miRNA826a/b, targeting *AOP2* transcript; light green: At4g03040 unknown function). The L*er*-0 sequence shows an inversion of *AOP2* and *AOP3* as well as an insertion (*AOP1.2*). The dominant SC leaf GSL is indicated on the right. (Modified figure according to[62]). **b** *GSL-AOP* main haplotype distribution. 1136 accessions were classified in four main *GSL-AOP* haplotypes and their frequency per ADMIXTURE population is shown (coloring according to main haplotypes). Accessions are distributed as follow: Admixed 137, Asia 79, Central Europe 184, Germany 171, Italy Balkan Caucasus 92, North Sweden 64, Relict 25, South Sweden 156, Spain 110 and West Europe 117. **c** *GSL-AOP* locus dendrogram of 74 accessions. Dissimilarity based Neighbor-Joining calculated dendrograms for the *GSL-AOP* locus (20kb, Chr4. 1.340.000-1.360.000). Colors show haplotype relationships based on sequence analysis of 1135 accessions. Red = H1, green = H2, blue = H3 and purple = H4. **d** *GSL-AOP* locus dendrogram and main haplotype groups for ONT sequenced accessions. Dissimilarity based Neighbor-Joining calculated dendrograms for *GSL-AOP*: 14.8kb Chr.4 pos. 12340082-1354901. Colored branches indicate haplotype groups. Col-0** represents the Reference sequence Chr.4 pos. 12340082-1354901 (CP002687), accessions from which additional amplicons were obtained are marked by (*), both amplicons are shown in the tree.

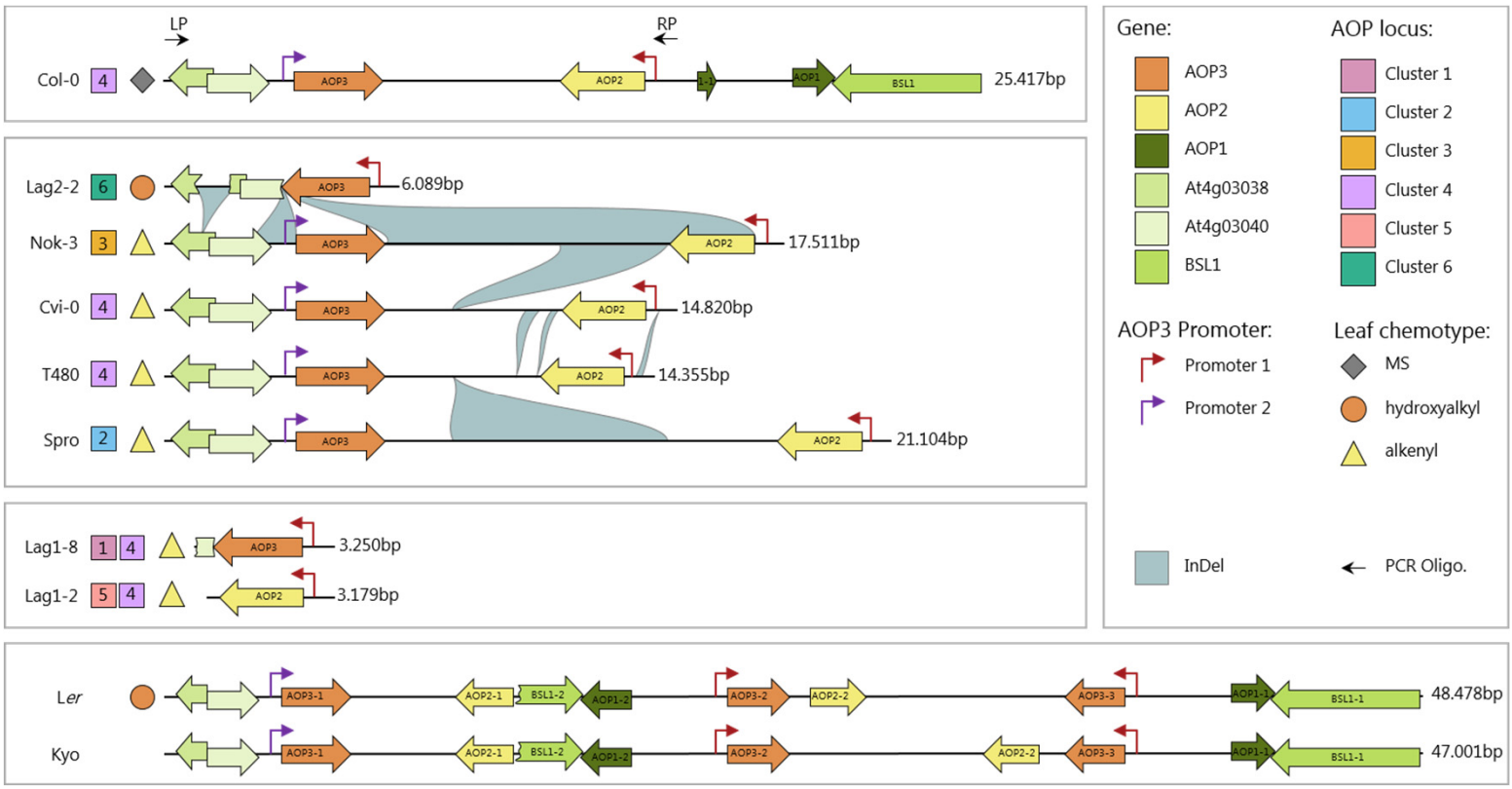

Figure 2

Synteny based diagram of *GSL-AOP* locus structure in accessions representing the main haplotypes found in this study. Colored arrows show combinations of genes: *AOP1* (dark green, unknown function), *AOP2* (yellow, alkenyl producing), *AOP3* (orange, hydroxyalkyl producing) and flanking genes (green: At4g03038, unknown function & miRNA826a/b, targeting *AOP2* transcript; light green: At4g03040 unknown function) as well as *BSL1* (green: *BRI1 SUPPRESSOR 1*). Symbols to the left indicate the dominant SC GSL in leaves and the *GSL-AOP* locus cluster according to ONT sequencing, length in bp is shown right to each locus. LP and RP indicate to position of the sequencing primers. First box from top: Col-0 Reference sequence Chr.4 pos. 12340082-1354901 (CP002687). Second box: Inversion and deletion events according to the new sequence information. Third box: Additional amplicons detected in the sequencing approach. Bottom: Ler and Kyo sequence according to Jiao and Schneeberger (2020).

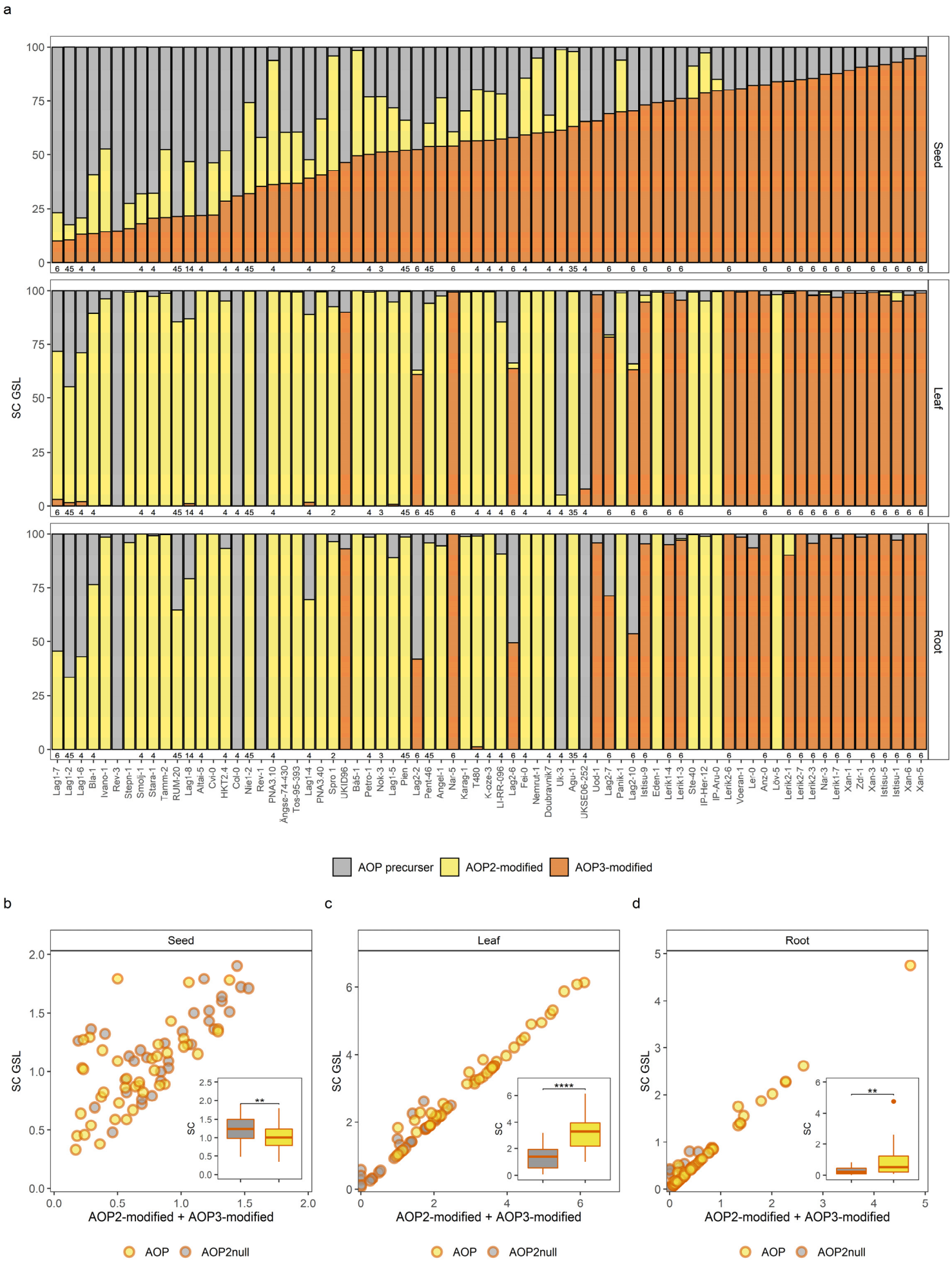

Figure 3 (see legend next page)

Figure 3

a Barplot of AOP-modified GSL distribution across tissues. The mean of the different groups of SC GSL in three different tissues seed, leaf and root, shown as percent of total SC GSLs. Accessions sorted by seed levels of AOP3-modified GSLs. Numbers indicate *GSL-AOP* haplotype cluster. Absolute GSL levels can be found in Extended Data Fig. 6. **b-d**. Scatterplot of AOP-modified GSL distribution across tissues. Means of the different groups of SC GSLs are shown in nmol/seed (seeds) or nmol/mg fresh weight (leaf, root). Accessions are coloured as grey/orange, AOP2null (accessions with no or only trace amounts (Istisu-1, Istisu-5, Istisu-9, Lag2-10, Lag2-2, Lag2-6, Lag2-7, Lerik2-1, Nar-3, Nar-5, Rev-1 and Uk-3) of AOP2-modfied GSLs in any of the tested tissues), or yellow/orange, AOP (accessions with AOP2-modified GSLs present in at least one tested tissue). AOP precursors, MT and MS GSLs; AOP2-modified GSLs, alkenyl and OH-alkenyl GSLs; AOP3-modified GSLs, (OH-alkyl and benzoyloxyalkyl GSLs). Boxplots of SC GSL by tissue and grouped by AOP and AOP2null accession. T test: ** Seed $p = 0.009$, ****Leaf $p > 0.0001$, ** Root $p = 0.0017$.

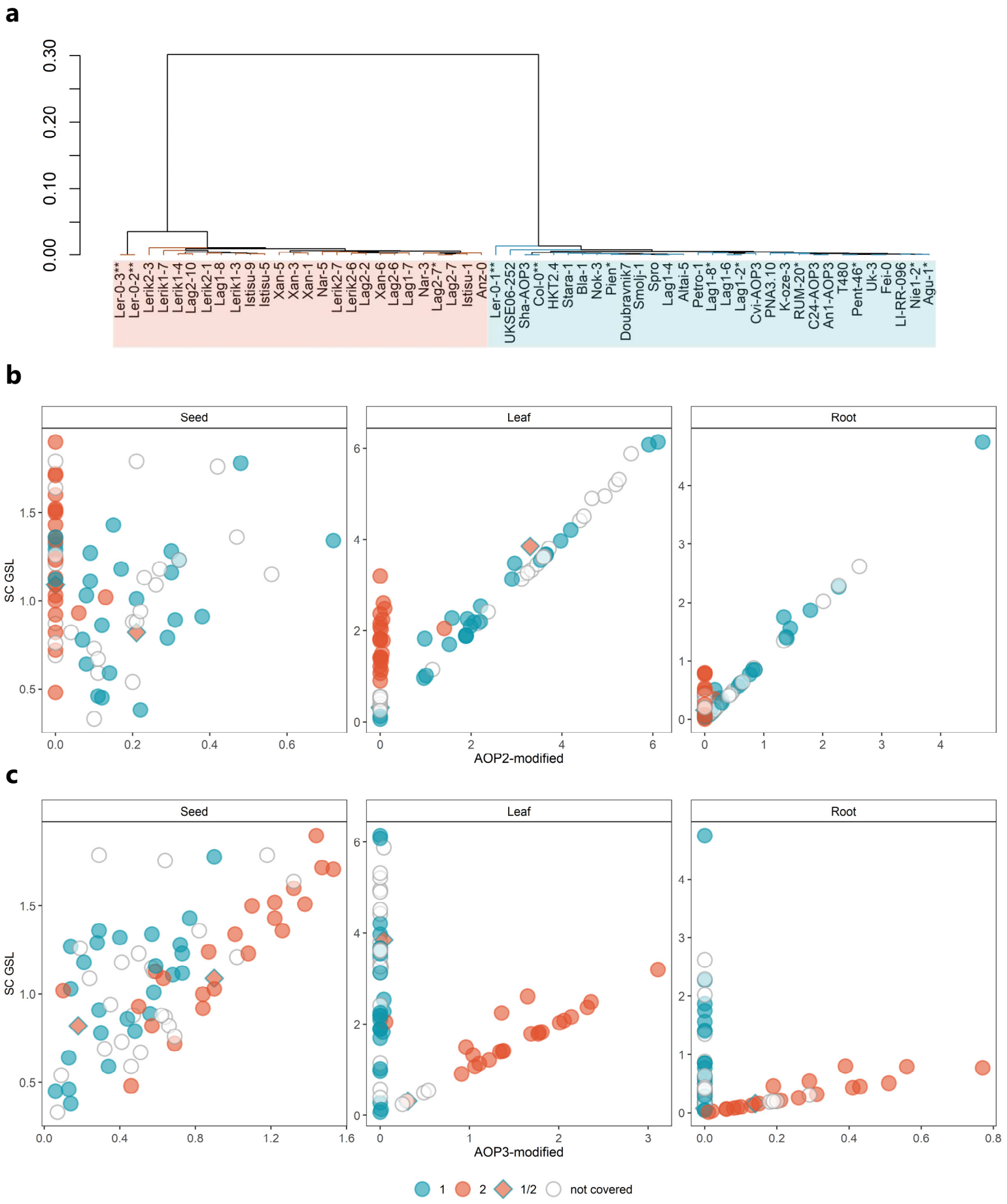

Figure 4 (see legend next page)

Figure 4

a Dendrogram and main *AOP3* promoters (1kb from ATG start codon) for ONT sequenced loci, including three sequences for L*er*-0 according to Jiao and Schneeberger (2020). Dissimilarity based neighbor-joining calculated dendrograms with colored branches indicating groups. Col-0** represents the reference sequence Chr.4 pos. 1353855-1354854 (CP002687), accessions from which additional amplicons were obtained are marked by (*), both amplicons are shown in the tree.
Scatterplot of **b** AOP2-modified and **b** AOP3-modified GSL distribution across tissues. The mean levels of for each accession are shown in nmol/seed (seeds) or nmol/mg fresh weight (leaf, root). Accessions are colored as white: no sequence information, blue: *AOP3* promoter cluster 1, red: *AOP3* promoter cluster 2 and diamond shape with yellow fill: accessions belonging to cluster 1 and 2.

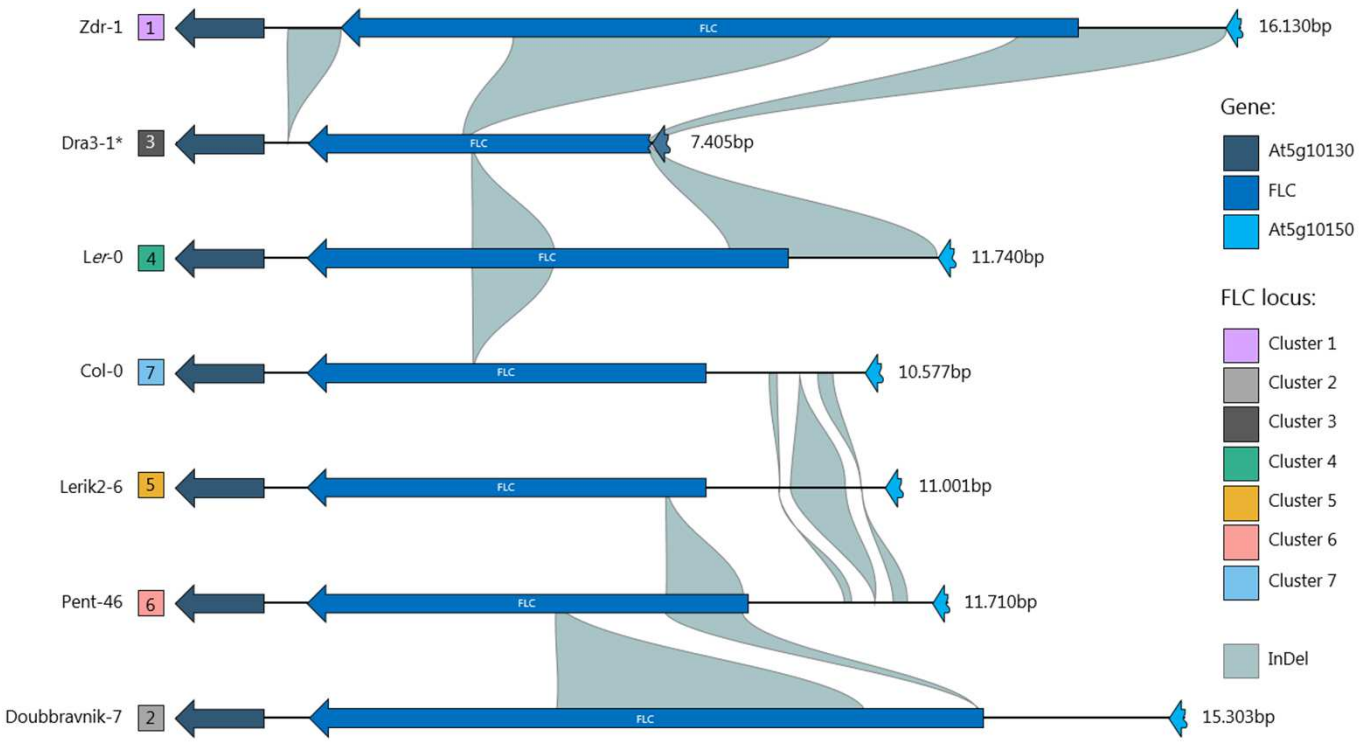

Figure 5
Synteny based diagram of *FLC* locus structure in accessions representing the main haplotypes found in this study. Colored arrows show combinations of genes: *FLC* (blue) and flanking genes *At5g10130* (dark blue), *At5g10150* (light blue). Colored squares to the left indicate the FLC locus cluster according to ONT sequences, length in bp is shown right to each locus. Shown is the amplified sequence according to Col-0 reference sequence Chr.5 pos. 3171361-3181999 (CP002687).

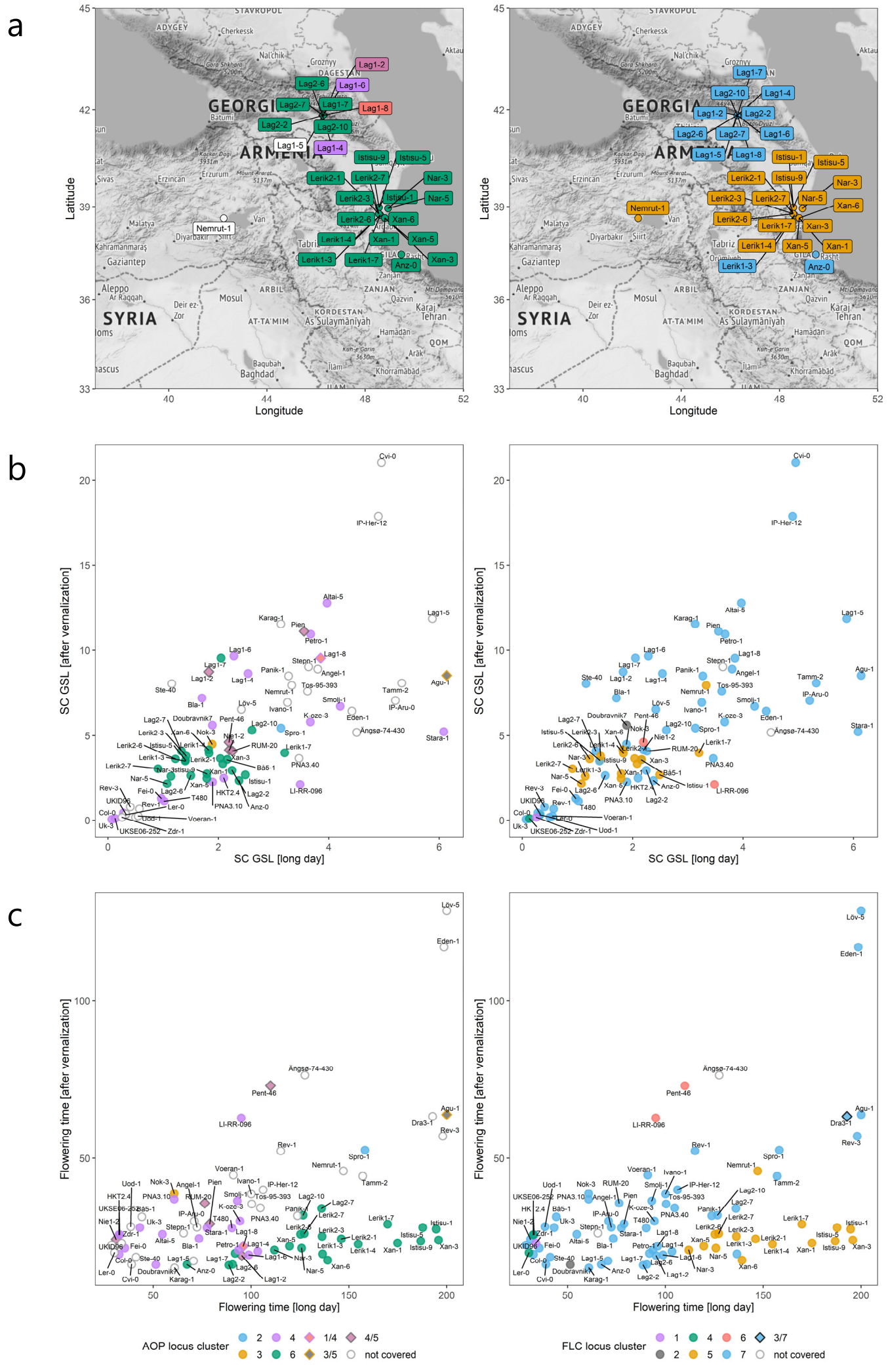

Figure 6 (see legend next page)

Figure 6

a Geographical distribution of Caucasus accessions. Colors indicate *GSL-AOP* locus cluster (left) and *FLC* locus cluster (right). Scatterplot of **b** leaf SC GSLs and **c** flowering time under LD conditions (x-axis) and after vernalization (y-axis). The mean SC-GSL for each accession is shown as nmol/mg leaf (fresh weight). Flowering time measured as days until first open flower under LD conditions or after 6 weeks of vernalization. Accessions are colored according to their *GSL-AOP* (left) and *FLC* (right) locus cluster determined by ONT sequencing. Diamond shaped locus symbols indicate accessions with more than one detected amplicon.

Figures

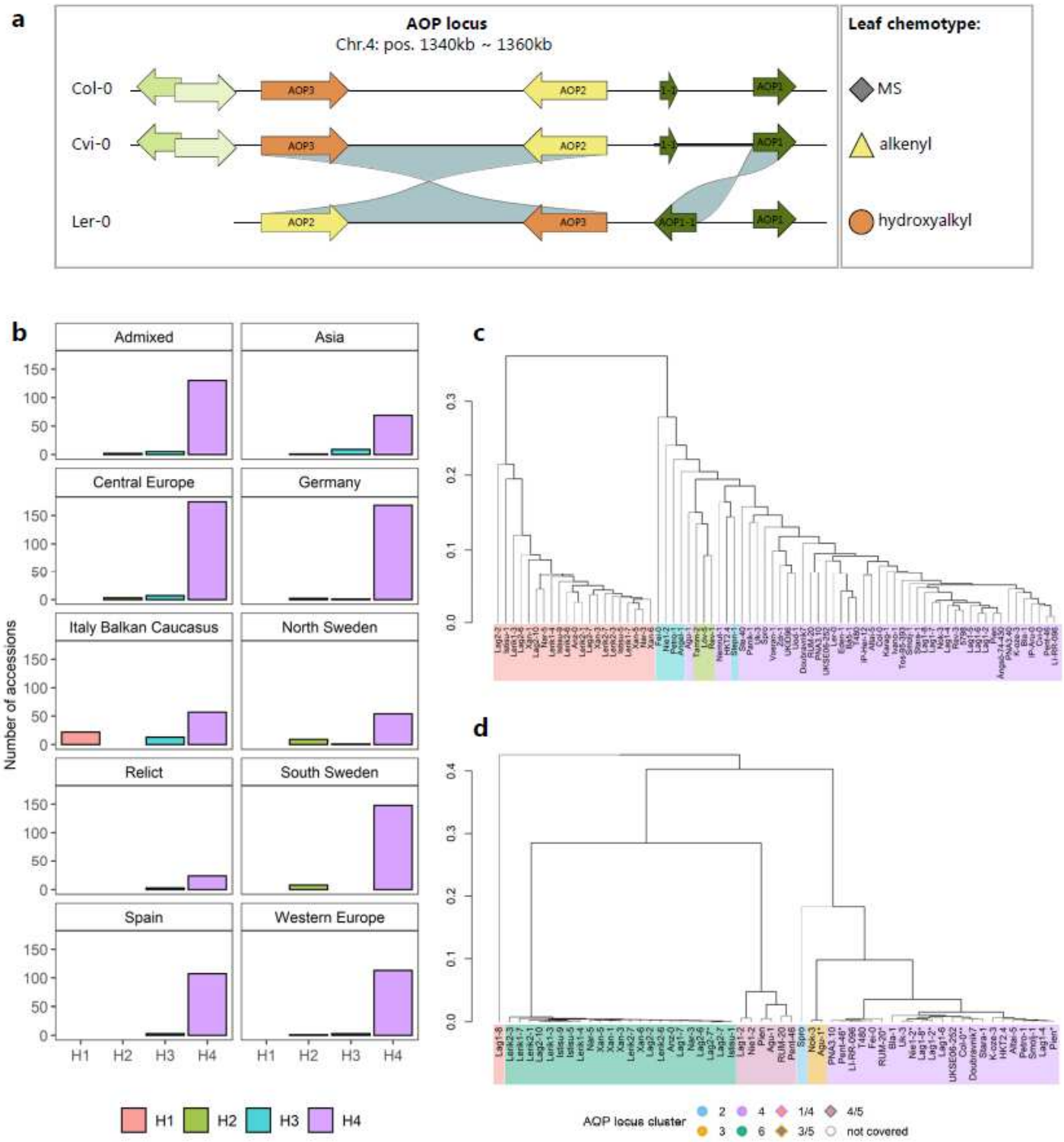

Figure 1

a GSL-AOP locus variation in A. thaliana determines the GSL leaf chemotype. Diagram of GSL-AOP locus structure in three accessions representing the main classical haplotype classes. Colored arrows show combinations of genes: AOP1 (dark green, unknown function), AOP2 (yellow, alkenyl producing), AOP3

(orange, hydroxyalkyl producing) and flanking genes (green: At4g03038, unknown function and miRNA826a/b, targeting AOP2 transcript; light green: At4g03040 unknown function). The Ler-0 sequence shows an inversion of AOP2 and AOP3 as well as an insertion (AOP1.2). The dominant SC leaf GSL is indicated on the right. (Modified figure according to62). b GSL-AOP main haplotype distribution. 1136 accessions were classified in four main GSL-AOP haplotypes and their frequency per ADMIXTURE population is shown (coloring according to main haplotypes). Accessions are distributed as follow: Admixed 137, Asia 79, Central Europe 184, Germany 171, Italy Balkan Caucasus 92, North Sweden 64, Relict 25, South Sweden 156, Spain 110 and West Europe 117. c GSL-AOP locus dendrogram of 74 accessions. Dissimilarity based Neighbor-Joining calculated dendrograms for the GSL-AOP locus (20kb, Chr4. 1.340.000-1.360.000). Colors show haplotype relationships based on sequence analysis of 1135 accessions. Red = H1, green = H2, blue = H3 and purple = H4. d GSL-AOP locus dendrogram and main haplotype groups for ONT sequenced accessions. Dissimilarity based Neighbor-Joining calculated dendrograms for GSL-AOP: 14.8kb Chr.4 pos. 12340082-1354901. Colored branches indicate haplotype groups. Col-0** represents the Reference sequence Chr.4 pos. 12340082-1354901 (CP002687), accessions from which additional amplicons were obtained are marked by (*), both amplicons are shown in the tree.

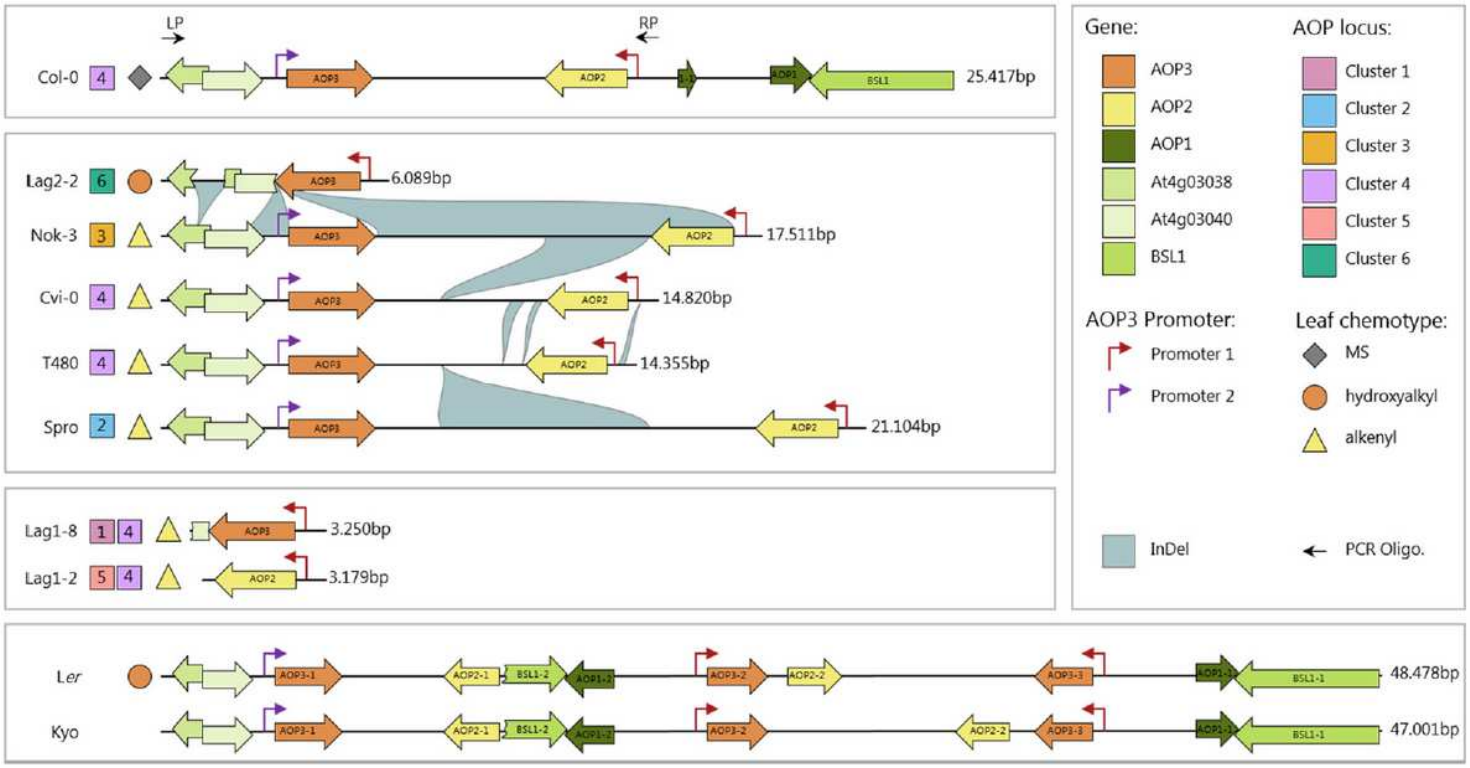

Figure 2

Synteny based diagram of GSL-AOP locus structure in accessions representing the main haplotypes found in this study. Colored arrows show combinations of genes: AOP1 (dark green, unknown function), AOP2 (yellow, alkenyl producing), AOP3 (orange, hydroxyalkyl producing) and flanking genes (green: At4g03038, unknown function & miRNA826a/b, targeting AOP2 transcript; light green: At4g03040 unknown function) as well as BSL1 (green: BRI1 SUPPRESSOR 1). Symbols to the left indicate the

dominant SC GSL in leaves and the GSL-AOP locus cluster according to ONT sequencing, length in bp is shown right to each locus. LP and RP indicate to position of the sequencing primers. First box from top: Col-0 Reference sequence Chr.4 pos. 12340082-1354901 (CP002687). Secondbox: Inversion and deletion events according to the new sequence information. Third box: Additional amplicons detected in the sequencing approach. Bottom: Ler andKyo sequence according to Jiao and Schneeberger (2020).

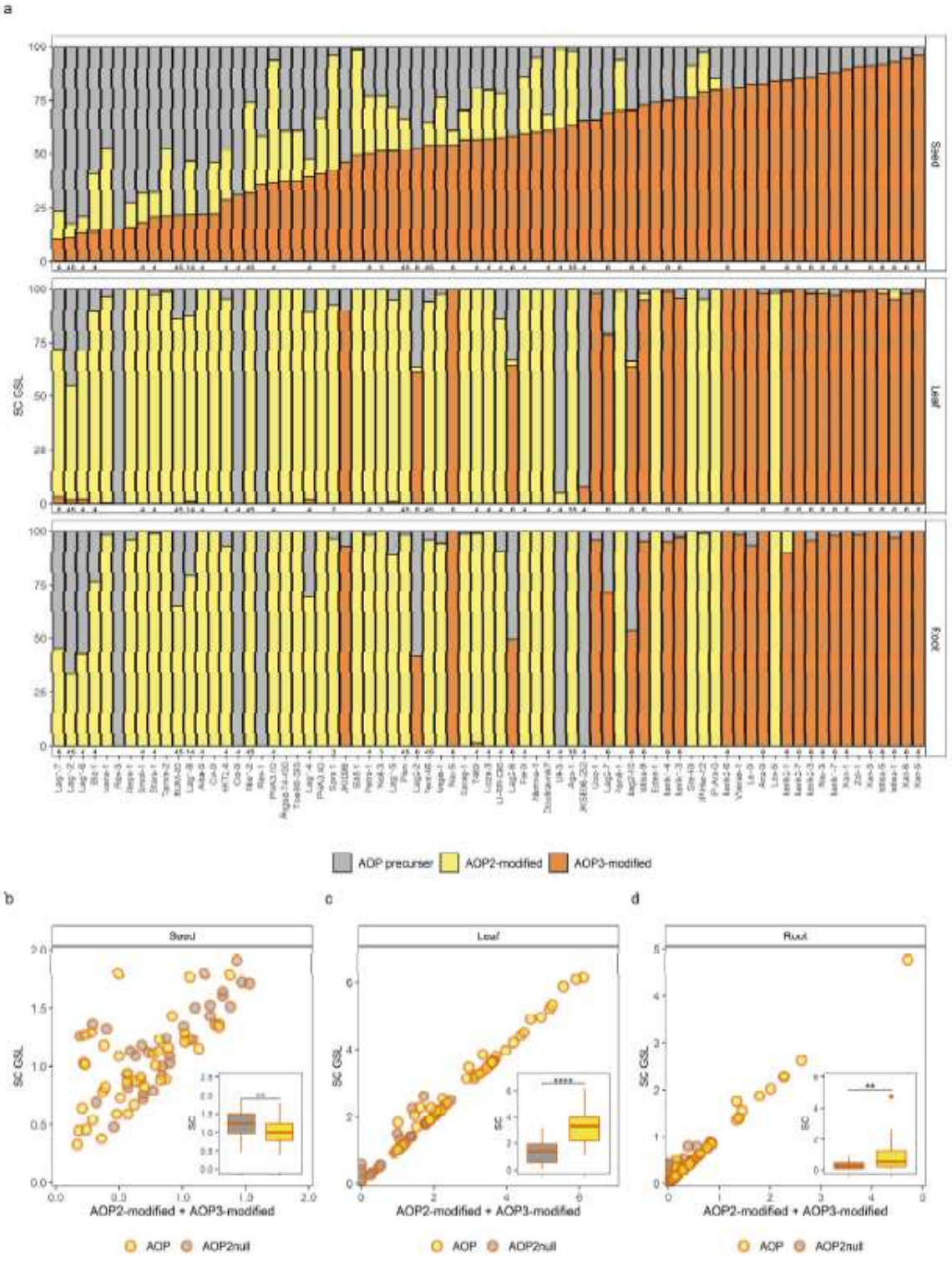

Figure 3

a Barplot of AOP-modified GSL distribution across tissues. The mean of the different groups of SC GSL in three different tissues seed, leaf and root, shown as percent of total SC GSLs. Accessions sorted by seed levels of AOP3-modified GSLs. Numbers indicate GSL-AOP haplotype cluster. Absolute GSL levels can be found in Extended Data Fig. 6. b-d. Scatterplot of AOP-modified GSL distribution across tissues. Means of the different groups of SC GSLs are shown in nmol/seed (seeds) or nmol/mg fresh weight (leaf, root). Accessions are coloured as grey/orange, AOP2null (accessions with no or only trace amounts (Istisu-1, Istisu-5, Istisu-9, Lag2-10, Lag2-2, Lag2-6, Lag2-7, Lerik2-1, Nar-3, Nar-5, Rev-1 and Uk-3) of AOP2-modfied GSLs in anyofthe tested tissues), or yellow/orange, AOP (accessions with AOP2-modified GSLs present in at least one tested tissue). AOP precursors, MT and MS GSLs; AOP2-modified GSLs, alkenyl and OH-alkenyl GSLs; AOP3-modified GSLs, (OH-alkyl and benzoyloxyalkyl GSLs). Boxplots of SC GSL by tissue and grouped by AOP and AOP2null accession. T test: ** Seed p = 0.009, ****Leaf p > 0.0001, ** Root p = 0.0017.

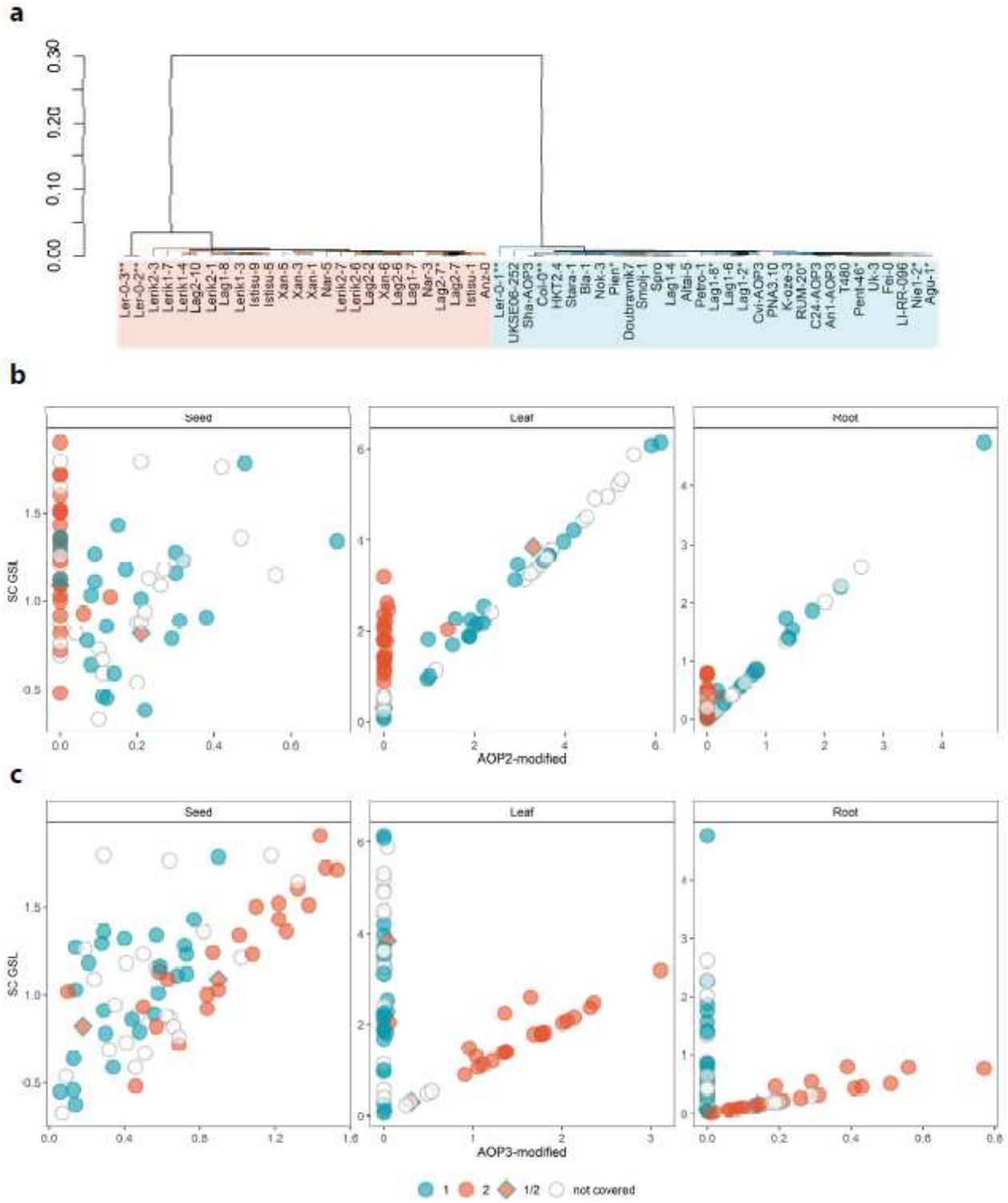

Figure 4

a Dendrogram and main AOP3 promoters (1kb from ATG start codon) for ONT sequenced loci, including three sequences for Ler-0accordingtoJiaoand Schneeberger (2020). Dissimilarity based neighbor-joining calculated dendrograms with colored branches indicating groups. Col-0** represents the reference sequence Chr.4 pos. 1353855-1354854 (CP002687), accessions from which additional amplicons were obtained are marked by (*), both amplicons are shown in the tree. Scatterplot of b AOP2-modified and b

AOP3-modified GSL distribution across tissues. The mean levels of for each accession are shown in nmol/seed (seeds) or nmol/mg fresh weight (leaf, root). Accessions are colored as white: no sequence information, blue: AOP3 promoter cluster 1, red: AOP3 promoter cluster 2 and diamond shape with yellow fill: accessions belonging to cluster 1 and 2.

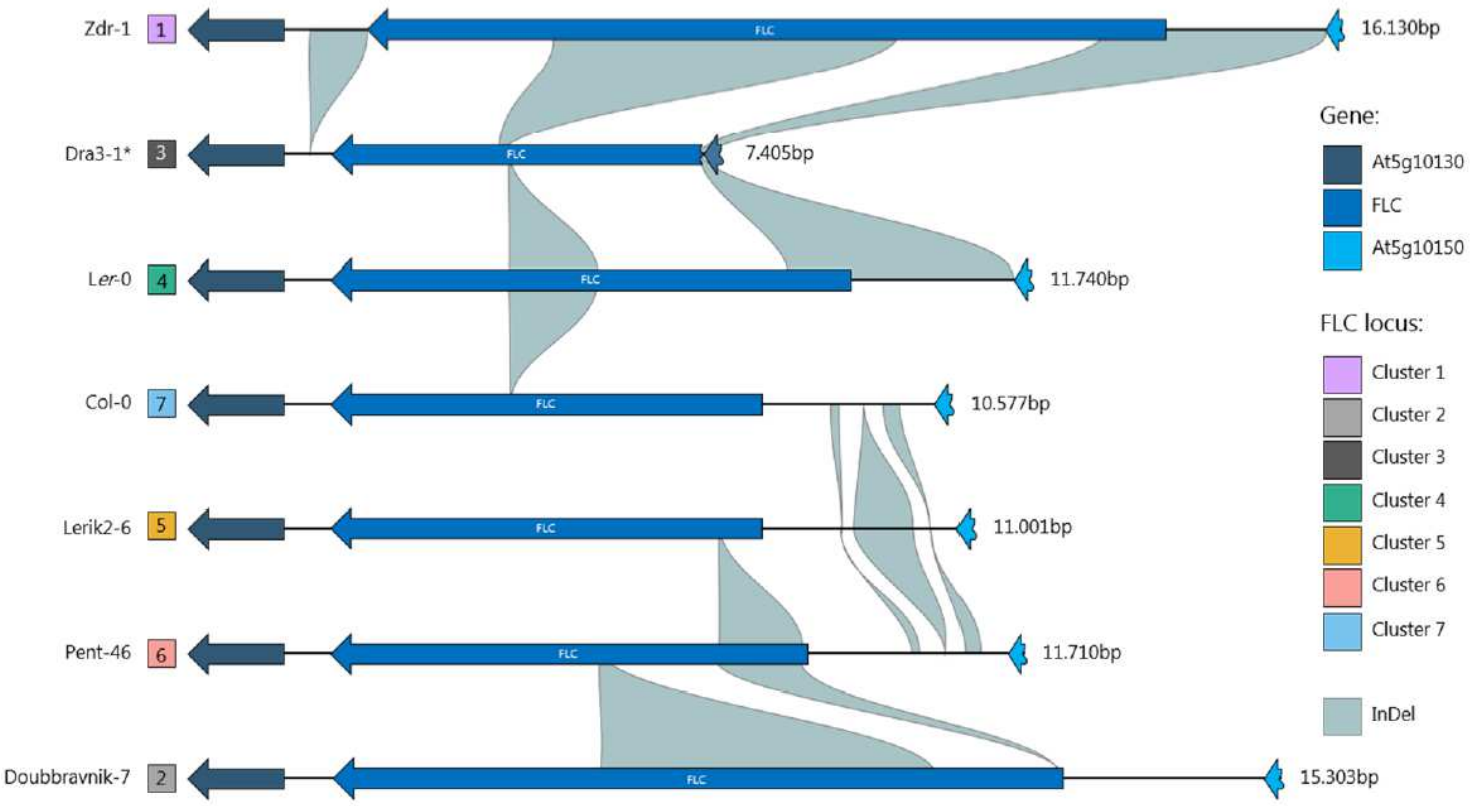

Figure 5

Synteny based diagram of FLC locus structure in accessions representing the main haplotypes found in this study. Colored arrows show combinations of genes: FLC (blue) and flanking genes At5g10130 (dark blue), At5g10150 (light blue). Colored squares to the left indicate the FLC locus cluster according to ONT sequences, length in bp is shown right to each locus. Shown is the amplified sequence according to Col-0 reference sequence Chr.5 pos. 3171361-3181999 (CP002687).

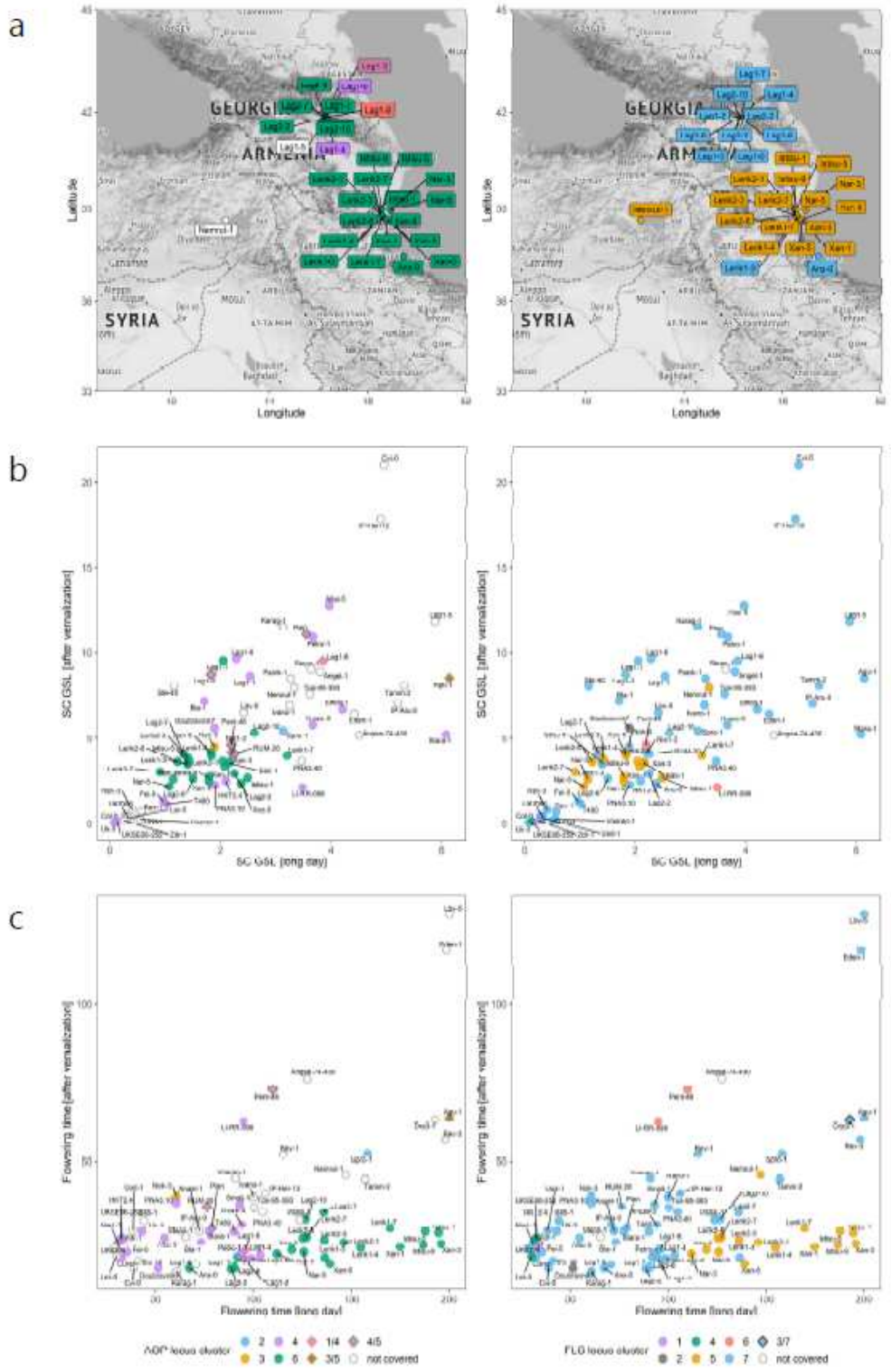

Figure 6

a Geographical distribution of Caucasus accessions. Colors indicate GSL-AOP locus cluster (left) and FLC locus cluster (right). Scatterplot of b leaf SC GSLs and c flowering time under LD conditions (x-axis) and after vernalization (y-axis). The mean SC-GSL for each accession is shown as nmol/mg leaf (fresh weight). Flowering time measured as days until first open flower under LD conditions or after 6 weeks of vernalization. Accessions are colored according to their GSL-AOP (left) and FLC (right) locus cluster

determined by ONT sequencing. Diamond shaped locus symbols indicate accessions with more than one detected amplicon.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ExtendedDataTab1.xlsx
- ExtendedDataTab2.xlsx
- ExtendedDataTab3.xlsx
- ExtendedDataTab4.xlsx
- Extendeddatafigures.pdf